META RESEARCH CONFERENCE

MERE 2019

PROCEEDINGS

DECEMBER 11, 2019

11:30H - 14:00H 55.309
14:00H - 15:00H 55.410

RESEARCH METHODOLOGY
UNIVERSITAT POMPEU FABRA
BARCELONA

Meta Research

Research Methods Course

Master in Sound and Music Computing, Master in Intelligent and Interactive Systems, Master in Computational Biomedical Engineering and Master in Wireless Communications

Information and Communication Technology Department
Universitat Pompeu Fabra, Barcelona

June 2020

Davinia Hernández-Leo
Judit Martínez-Moreno
(Eds.)

Editors

Davinia Hernández-Leo
Associate Professor, Serra Húnter Fellow
Universitat Pompeu Fabra, Barcelona
E-mail: davinia.hernandez-leo@upf.edu

Judit Martínez-Moreno
PhD Student, Teaching Assistant
Universitat Pompeu Fabra, Barcelona
E-mail: judit.martinez@upf.edu

Universitat Pompeu Fabra, Barcelona
e-repository UPF, http://repositori.upf.edu/

Preface

This document collects a selection of papers written by master's students in the context of the "Research Methods" course common to the Master's Programmes in Sound and Music Computing, Intelligent and Interactive Systems, Computational Biomedical Engineering and Wireless Communications, of the Information and Communication Technology Department at Universitat Pompeu Fabra, Barcelona, during the 2019-2020 academic year.

The papers were written as part of an integrative assignment entitled "Meta-Research", where students were expected to do a small piece of research about a transversal research topic. Students worked in teams and selected a topic, among the following suggested themes:

- Biases in technology research
- Research methods and evaluation of technology
- Responsible research and public engagement
- Doctoral studies

A refinement of the topic, the particular research questions to study and the methodology to apply were proposed by the students and discussed with the course educators in tutoring sessions. A total of 13 papers were written by the students and presented in the classroom. Assessment included peer-review by students during the presentations, through a conference management program and assessment by the educators.

Table of contents

Sources of bias in research linking race to intelligence

Master in Intelligent and Interactive Systems

M. Acera[1], A. Aguado[2], A. Moral[3], and C. Morales[4]

[1] mario.acera01@estudiant.upf.edu
[2] alfonso.aguado01@estudiant.upf.edu
[3] albert.lleo3@gmail.com
[4] cristian.morales01@estudiant.upf.edu

Abstract. The relationship between race and intelligence has always been the cause of great controversy, and in which bias plays an impor- tant role. The purpose of this study is to provide a summary of the main sources of bias in studies linking race to intelligence in order to help avoid them in future research. Through our literature review we concluded five main sources of error: Conflicts of Interest, Sample bias, Control for Relevant Variables, Correlation as Causation and Use of Ge- netics. The latter appears as a new tool for justifying racial differences in cognitive performance, making use of the recent mapping of the human genome and the subsequent studies identifying unique genetic markers for different races.

Keywords: race and intelligence · intelligence heritability · environmental and genetic influences · IQ racial gap · scientific racism.

1 Introduction

It is believed that the study of cognitive differences between individuals started four thousand years ago in what is now China [1]. For three thousand years, the chinese emperors employed a system of exams to select the best of- ficials who would protect them. But the scientific study of differences between individuals, carried out today by a branch of psychology called Differential Psy- chology, started at the end of the 19th century.

The second half of the 19th century was a breeding ground for the develop- ment of a scientific psychology of individual differences: in 1859 Charles Darwin published "On the Origin of Species by means of Natural selection" [2]. His cousin Francis Galton, inspired by his reading of Darwin, started to be inter- ested in the application of the laws of biological science to the improvement of the qualities of human beings [3]. With this purpose, Galton created in London, in the beginning of the 20th century, a program to artificially produce a better human race through regulating marriage and thus procreation. This program, known as the eugenics movement, began by encouraging intelligent people to have more children than less intelligent people. Soon, as a result of this move- ment, several states of the United States of America enacted laws [4] to force

the sterilization of the less educated and of minority populations. It caused the sterilization of more than sixty thousand people in the United States. But the most terrible crimes in the name of the eugenics movement were committed by the Nazi Germany during the World War II [5].

The first tests of intelligence as we know them today were developed in 1905 by the french Alfred Binet and Theodore Simon when they both developed "The Metric Scale of Intelligence" [6]. In contrast to Galton's work, Binet and Simon wanted to help children with developmental delays to develop their cognitive skills. A decade later, in 1916, Lewis M. Terman [7] would develop the intelli- gence quotient (IQ) and the best-known test of intelligence: the Stanford-Binet Revision. The tests of intelligence applied nowadays are based on the work of Binet, Simon and Terman.

In contrast to how Binet and its colleagues in Europe applied the intelligence tests, Robert M. Yerkes [8] who believed that there were genetic differences in intelligence between races, employed the Army Alpha and Beta tests of intelli- gence during World War I to classify soldiers according to their mental ability and to identify those who are weak-minded.

After the terrible crimes carried out by the Nazi Germany during the World War II, the scientific publications that linked race and intelligence were non-existent or very limited. But in 1969, the Berkeley emeritus professor of psychol- ogy Arthur R. Jensen published an article [9] in the Harvard Education Review that concluded that gaps in intelligence test results between black and white students might be because of genetics. This article arose an intense debate due to the fact that, at that time, there was no evidence to support Jensen claims. He just concluded that scientists would discover the "intelligence genes" in the future, without no conclusive data to support such argument.

In 1984, James R. Flynn published the article "The mean IQ of Americans: Massive gains 1932 to 1978" [10] where he demonstrated that americans had gained 13.8 points of IQ over a period of 46 years. These results suggested that environmental factors such as education or health care play an important role in developing intelligence, in contrast to what it was believed by those who see intelligence as an innate trait.

Ten years after the publication of the Flynn's article, Charles Murray and Richard J. Herrnstein published what has been become a reference book [11] for those who support the idea of intelligence as an innate trait and to those who be- lieve that gaps in intelligence between races exist. "The Bell Curve: Intelligence and Class Structure in American Life" has sold hundreds of thousands copies since 1994 and still does, being today one of the most sold books in Amazon USA. Additionally, it has been cited by other academic publications more than ten thousand times.

The main goal for this study is to identify common sources of bias and main pitfalls in this area of research. Literature reviews usually center in reporting the main studies in the area and its potential biases but do not group them into general sources of bias as presented in this study. Furthermore, studies linking race to intelligence can have a strong impact on the life of millions of people, especially to those that belong to minority populations. All studies, but par- ticularly those who have a direct effect in the lives of vast amounts of people, should be very rigorous in their methodologies and in suggesting any results or conclusions. We expect to help researchers by finding categories of bias so they can be used as a reference for identifying them when citing or while conducting new studies

It should be noted that we are not concerned in this study with definitional problems associated with the constructs of intelligence or the categorization of human genetic diversity into distinct races.

2 Research Methodology

In order to achieve the goal of this research, a review method has been ap- plied to collect, analyze and compare results, ideas and conclusions that have been exposed in the target field.

Since it has been shown that there are a lot of contradictions and different ideas and controversies about this concrete topic, it was decided to conduct a literature review to assess the current state of the field. Although the purpose is trying to be as rigorous as possible, there is always a possibility of excluding some relevant research pieces given the scope of this project, but this could be palliated by a deep research of the resources available in future research.

Google Scholar, the Web of Science and Mendeley were used in order to find studies on race and intelligence, to identify, select and document those that fall into systematic errors. The keywords used were:“race and intelligence”, “intelli- gence heritability”, “environmental and genetic influences”, “group differences in intelligence”, “IQ racial gap” and “improvement in IQ performance”, “scientific racism”. The Google search engine was also used to seek further information and additional facts for concrete cases. The criteria for taking into consideration the papers were how relevant they were about the subject. On this selection criteria the number of citations and the importance of the journal in which they were published has been taken into account and whether we can found identifiable sources of bias. Additionally, it has been proven useful for us searching for the most relevant papers that cited the one we found. It is a useful way to find new papers and authors in the field.

3 Results

In this section we will show the main categories of bias that we have found when analysing the principal studies on racial differences in intelligence that have been published since the second half of the 20th century. Table 1 shows the studies included in this section and which categories are relevant to each of them.

3.1 Conflicts of interest

Conflicts of interest increase the likelihood of biases. It can involve research sponsors, journals, publishers or authors, among others. For instance, the con- servative Bradley Foundation helped fund [12] the research of The Bell Curve bestseller book in 1994 that we have described in the Introduction section. Ad- ditionally, Charles Murray, who co-authored the book with Richard J. Herrn- stein, has been a fellow of one of the most influential conservative think tanks since 1990, the American Enterprise Institute. Another example are the stud- ies published in the Mankind Quarterly journal. According to a report in the Independent newspaper, the institute received USD 50,000 in 1993 from the Pi- oneer Fund known as one of the most controversial foundations promoting racial discrimination[13]. This organization is known for funding multiple controversial publications, including The Bell curve and The Minnesota study [14]. Therefore, it is important to check for possible conflicts of interests when evaluating the validity of a certain study and journal.

3.2 Sample bias

In this category we also included data manipulation. For example, last edi- tor in chief of the journal Mankind Quarterly, Gerhard Meisenberg, published an article[15] in 2010 on the relationship between intelligence and economic success. The conclusion of the study was carried out by using data that was manipulated, as Jelte M. Wicherts pointed out[16] one year before. Samples of Africans with the highest IQ scores were deliberately excluded from the study. As a conse- quence of this, the average IQ score of the representative sample of the study was lower than it should be. Another important case of this bias is the Minnesota Transracial Adoption Study [14] which stated racial differences in IQ but failed to mention that an on average the African American adoptees had a higher age of adoption compared to that of the White adoptees.

3.3 Control for relevant variables

As previously discussed, Jensen stated that innate differences in intelligence were the cause of poor performance in school for black children[9]. However, there is now a general consensus that there is a strong relationship in the opposite di- rection, that is, education is positively influences IQ, especially in early childhood

[17]. This same source of bias is also present in The Bell Curve [11]. Moreover, a 1997 study showed that a year of schooling will increase performance on the test used in the analysis presented by this book [18]. This is an important con- clusion given that in the US access to good education is highly dependent on race.

Other studies have shown that performance in African American students is hindered when test takers believe that they are part of a group known for performing poorly. This is referred to as the "stereotype threat", in which ex- aminees can perform better or worse depending on what group they believe they belong to (good or bad performers). This is thought to be a motivation variable, as opposed to a cognitive one [19] [20].

3.4 Correlation as causation

Although it is a well-known scientific principle, it has been a general prob- lem in studies relating intelligence to race or countries. For instance, a popular theory explaining cognitive differences across races, is that populations faced with harder climatic conditions had undergone more severe selection conditions favoring intelligence. This resulted in people with higher cognitive abilities [21] [22]. In this case the authors only considered one possible explanation for the seen differences in IQ: climate. As seen in the studies aforementioned, it is not sensible simplifying IQ to only one variable.

In another study by Rushton, the authors concluded that Whites have larger brains than Blacks in the U.S. This study was based on exterior skull mea- surements [21]. In later studies the authors reaffirm that there is a 15-18 point average IQ gap between Black and White Americans, and the main reason for the gap lies in a brain size difference[23] [24]. Even if these measurements are correct, concluding that bigger skulls or bigger brains imply more intelligence is not self-evident and needs to be proven. As pointed by Hunt in 2007, intelligence is an extremely complex topic, influenced by multiple variables[25]. The influ- ence of such variables is not only difficult to estimate, but also highly variable across populations.

3.5 Use of genetics

This bias type might be thought as a particular case of correlation as cau-sation but we believe is worth isolating. An example of the use of genetics is the study by Evans et al, which mapped the distribution of variations of the gene Microcephalin (MCPH1) in the world. This gene has been proven to regulate brain size, and a recent mutation, referred to as haplogroup D, evolved under strong positive selection in the human evolutionary lineage. *Figure 1* shows the distribution of haplogroup D in the globe [26].

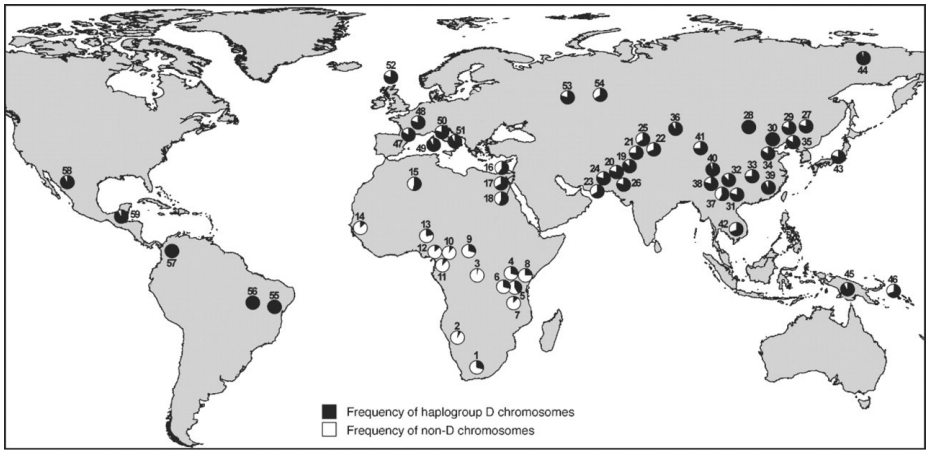


Fig. 1. Frequency of haplogroup D chromosomes. Adapted from "Evans, P. D., Gilbert, S. L., Mekel-Bobrov, N., Vallender, E. J., Anderson, J. R., Vaez-Azizi, L. M.,Lahn, B. T. (2005). Microcephalin, a gene regulating brain size, continues to evolve adaptively in humans. Science, 309(5741), 1717–1720."

The authors hypothesize that D and non-D haplotypes have different effects structuring the brain, arguing recent evolution in human brains. Although the study does not make explicit claims of race, in a subsequent paper the authors state that "data offer strong evidence that haplogroup D emerged very recently and subsequently rose to high frequency under strong positive selection" [27]. A similar pattern has been shown in another gene, ASPM, which is also involved in the determination of brain size [28]. Posterior studies have explained that positive selection might not be the only explanation for the expansion of such genes, in which randomness and mutations that happen at the edge of an ex- panding population play a role [29][30]. Furthermore, some recent genetic studies controlling for multiple environmental factors have discovered only a mild gap in intelligence in African descendant population in the US [31]. These are also examples of correlation as causation.

Another example of the misuse of genetics is the 2014 book by Nicholas Wade "A Troublesome Inheritance: Genes, Race, and Human History" in which the author uses genetic differences across populations as a possible explanation for intelligence differences. However, genetics is a recent science in which the mecha- nisms of how genes express themselves or how they spread across populations are still unclear. It is important to note that this book has been widely discredited by the same geneticists whose data were used for this book[32]. In such delicate matters, it is crucial to acknowledge the inherent uncertainty of the methods used.

Table 1. List of publications analyzed

Author	Type	SB	CI	CRV	CC	UG	Citations	Journal	Journal Impact
C. Murray,RJ Herrnstein, 1994 [11]	book	*	*	*	*		> 10500		
A. Jensen, 1969 [9]	paper			*			> 5000	Harvard Educational Review	2.190
R. Lynn, M. Stuart, 2002 [33]	book			*			> 900		
J. P. Rushton, 1995 [21]	book				*		> 850		
R. Lynn, 2006 [34]	book		*		*		> 300		
S. Kanazawa, 2008 [35]	paper	*					> 100	Intelligence	3.274
G. Cochran et al., 2006 [36]	paper			*	*		> 200	Journal of Biosocial Science	702
R. Lynn, G. Meisenberg, 2010 [15]	paper	*					> 150	Intelligence	3.168
R. Lynn, 1991 [22]	paper		*		*		> 100	Mankind Quarterly	1.1(2010)
J. Rushton, A. Jensen, 2005 [23] [24]	paper				*		> 50	Psychology, Public Policy, and Law	0.69
D. Templer, 2008 [37]	paper	*					> 50	Personality and Individual Differences	1.598
J. Rushton, A. Jensen, 2010 [24]	paper				*		> 50	The Open Psychology Journal	N/A
K. Eyferth, 1961 [38]	paper	*					> 70	Archiv fur die gesamte Psychologie	N/A
R. A. Weinberg, S. Scarr 1992 [14]	paper	*	*	*			> 150	Intelligence	4.388(2018)
P. D. Evans et al., 2005 [26]	paper					*	> 600	Science	30.927
N. Mekel-Bobrov et al, 2005 [27]	paper					*	> 500	Science	30.927
B. F. Voight,S. Kudaravalli [28], 2006 [27]	paper					*	> 2000	PLoS Biology	14.101
N. Wade, 2015 [39]	book					*	> 250		

SB - Samble Bias
CI - Conflict of Interest
CRV - Control of Relevant Variables
CC - Caurrelation as Causation
UG - Use of Genetics

4 Conclusions

This study provided a review of the main errors and biases found in the existing literature linking intelligence to race. We believe that understanding the most common sources of errors could help academic journals and authors to improve the quality of the studies they publish by correctly assessing the validity of their citations and by encouraging them to be more diligent when conducting studies of this kind. In this regard, this article could be understood as a guideline for common conceptual mistakes to avoid in order to produce more responsible articles on intelligence and race.

By using a structured methodology we have analysed eighteen studies on intelligence and race. As a result of it, we have identified five recurrent errors that are found in many of these studies: *Conflicts of Interest*, *Sample Bias*, *Con- trol for Relevant Variables*, *Correlation as Causation* and *Use of Genetics*. It is worth mentioning that these types of studies could directly affect the lives of millions of people as it happened when the eugenics movement was executed in the United States and in the Nazi Germany. Furthermore, this line of research can also have an indirect impact on the population when our policy makers are influenced by their results and conclusions. Therefore, it is extremely important to be aware of the limitations and biases of this area of research.

One limitation of this study is that it only focused on the most common er- rors and in the most cited papers, thus, potentially excluding important studies. On the other hand, our study is restricted to the academic context but many other publications happen outside scientific journals, so it might be relevant to include information published in magazines or websites.

It should be noted that genetics seem to be causing a new wave of studies explaining cognitive differences in different human races. We are aware that a more comprehensive literature review is needed in order to assert this. How- ever, since the completion of the Human Genome Project in 2003, there seems to be a tendency of mapping intelligence to DNA sequences. As stated before, the variables that control intelligence are various, and as such, finding spurious correlations between intelligence and genes could be used by those looking to explain the IQ gap between races.

References

1. A. Sanchez-Elvira Paniagua, P. J. Amor Andres, E. Fernandez Jimenez, and M. Olmedo Montes, *Introduccion al estudio de las diferencias individuales*. Sanz y Torres, 2005.
2. C. Darwin, "The Origin of Species, by Means of Natural Selection, or the Preservation of Favored Races in the Struggle for Life," *The Crayon*, 1860.
3. D. Aubert-Marson, "Sir Francis Galton: The father of eugenics," *Medecine/Sciences*, 2009.
4. G. Rodgers, "Yin and yang: the eugenic policies of the United States and China: is the analysis that black and white?," *Houston journal of international law*, 1999.
5. H. K. Anheier and S. Kuhl, "The Nazi Connection: Eugenics, American Racism, and German National Socialism.," *Contemporary Sociology*, 1995.
6. N. Jelliffe, "The Development of Intelligence in Children (The Binet-Simon Scale)," *The Journal of Nervous and Mental Disease*, 1917.
7. C. Karier and H. L. Minton, "Lewis M. Terman: Pioneer in Psychological Testing," *The Journal of American History*, 1990.
8. Carson J., "Army Alpha, Army Brass, and the Search for Army Intelligence," *Isis*, 1993.
9. A. Jensen, "How Much Can We Boost IQ and Scholastic Achievement," *Harvard Educational Review*, 1969.
10. J. R. Flynn, "The mean IQ of Americans: Massive gains 1932 to 1978," *Psychological Bulletin*, 1984.
11. C. Herrnstein, C. Aubrey, and R. J. Murray, "The Bell Curve: Intelligence and Class Structure in American Life," *British Journal of Educational Studies*, 1995.
12. B. Miner, "Who Is Backing "The Bell Curve"?.," *Educational Leadership*, 1995.
13. A. Saini, *Superior: The Return of Race Science*. Beacon Press, 2019.
14. R. A. Weinberg, S. Scarr, and I. D. Waldman, "The minnesota transracial adoption study: A follow-up of iq test performance at adolescence," *Intelligence*, vol. 16, no. 1, pp. 117–135, 1992.
15. R. Lynn and G. Meisenberg, "National IQs calculated and validated for 108 na- tions," 2010.
16. J. M. Wicherts, C. V. Dolan, and H. L. van der Maas, "The dangers of unsystematic selection methods and the representativeness of 46 samples of African test-takers," 2010.
17. R. E. Snow, "Cognitive-conative aptitude interactions in learning.," in *Abilities, motivation, and methodology: The Minnesota Symposium on Learning and Indi- vidual Differences.*, 1989.
18. C. Winship and S. Korenman, "Does Staying in School Make You Smarter? The Effect of Education on IQ in The Bell Curve," in *Intelligence, Genes, and Success*, pp. 215–234, Springer New York, 1997.
19. C. M. Steele and J. Aronson, "Stereotype Threat and the Intellectual Test Performance of African Americans," *Journal of Personality and Social Psychology*, vol. 69, no. 5, pp. 797–811, 1995.
20. C. Steele and J. Aronson, "Stereotype threat and the test performance of academ- ically successful African Americans.," 1998.
21. J. P. Rushton, *Race, evolution, and behavior : a life history perspective*. Transaction Publishers, 1995.
22. R. Lynn, "Race differences in intelligence : a global perspective," *Mankind Quaterly*, 1991.

23. J. P. Rushton and A. R. Jensen, "Wanted: More race realism, less moralistic fallacy," *Psychology, Public Policy, and Law*, 2005.
24. J. P. Rushton and A. R. Jensen, "Race and IQ: A Theory-Based Review of the Research in Richard Nisbett s Intelligence and How to Get It," *The Open Psychology Journal*, 2010.
25. E. Hunt and J. Carlson, "Considerations Relating to the Study of Group Differences in Intelligence," *Perspectives on Psychological Science*, vol. 2, pp. 194–213, jun 2007.
26. P. D. Evans, S. L. Gilbert, N. Mekel-Bobrov, E. J. Vallender, J. R. Anderson, L. M. Vaez-Azizi, S. A. Tishkoff, R. R. Hudson, and B. T. Lahn, "Microcephalin, a gene regulating brain size, continues to evolve adaptively in humans," *Science*, 2005.
27. N. Mekel-Bobrov, S. L. Gilbert, P. D. Evans, E. J. Vallender, J. R. Anderson, R. R. Hudson, S. A. Tishkoff, and B. T. Lahn, "Ongoing adaptive evolution of ASPM, a brain size determinant in Homo sapiens," *Science*, 2005.
28. B. F. Voight, S. Kudaravalli, X. Wen, and J. K. Pritchard, "A map of recent positive selection in the human genome," *PLoS Biology*, 2006.
29. S. Klopfstein, M. Currat, and L. Excoffier, "The fate of mutations surfing on the wave of a range expansion," *Molecular Biology and Evolution*, 2006.
30. L. Excoffier and N. Ray, "Surfing during population expansions promotes genetic revolutions and structuration," jul 2008.
31. R. E. Nisbett, J. Aronson, C. Blair, W. Dickens, J. Flynn, D. F. Halpern, and E. Turkheimer, "Intelligence: New Findings and Theoretical Developments," *Amer- ican Psychologist*, 2012.
32. H. M., "Racism, the misuse of genetics and a huge scientific protest - Los Ange- les Times. 12 Aug. 2014. https://www.latimes.com/business/hiltzik/la-fi-mh-huge- scientific-protest-20140812-column.htmlpage=1. 10 Nov. 2019 ."
33. R. Lynn, T. Vanhanen, and M. Stuart, *IQ and the wealth of nations*. Greenwood Publishing Group, 2002.
34. R. Lynn, *Race differences in intelligence: an evolutionary analysis.* Washington Summit Publishers, 2006.
35. S. Kanazawa, "Temperature and evolutionary novelty as forces behind the evolu- tion of general intelligence," *Intelligence*, vol. 36, no. 2, pp. 99–108, 2008.
36. G. Cochran, J. Hardy, and H. Harpending, "Natural history of Ashkenazi intelligence," *Journal of Biosocial Science*, 2006.
37. D. I. Templer, "Correlational and factor analytic support for rushton's differential k life history theory," *Personality and Individual Differences*, vol. 45, no. 6, pp. 440– 444, 2008.
38. K. Eyferth, "Performance of different groups of children of occupation forces on the hamburg—wechsler intelligence test for children (hawik)," *Archiv für diegesamte Psychologie*, pp. 113–222, 1961.
39. N. Wade, *A troublesome inheritance: Genes, race and human history.* Penguin, 2015.

Study of publication bias in antidepressant clinical trials

Courtney Rose Belin[1], Marta Borràs Argemí[2], Mariona Forcada Romeu[3], Adrià Mas Dalmases[4]

Master in Computational Biomedical Engineering
Universitat Pompeu Fabra (UPF), Barcelona, Spain

courtneyrose.belin01@estudiant.upf.edu[1]
marta.borras02@estudiant.upf.edu[2]
mariona.forcada01@estudiant.upf.edu[3]
adria.mas02@estudiant.upf.edu[4]

Abstract. Publication bias appears in many different cases of research and clinical trials as a result of the tendency to publish manuscripts based on the direction or strength of the study findings. Many clinical trials evaluating the efficacy of antidepressant drugs have been performed but the publication biases have not been extensively analyzed. This article compares the differences between published and unpublished depression clinical trials from the *clinicaltrials.gov* website between the dates of 01/01/2013 to 01/01/2018 by focusing on the results' success. The research was conducted so that it is straightforward to conclude a significant relation between the probability of publishing an article in a medical journal and the positiveness of the results obtained during the medical trials. In addition, a previous study conducted by a different group offers insight into whether publication bias in depression clinical trial research has changed over the past decades.

Keywords: Depression, Antidepressant Drugs, Clinical Trials, Publication Bias

1 Introduction

Bias is defined as the deviation of results or inferences from the truth. Any trend in the collection, analysis, publication, interpretation or review of data that leads to conclusions that are systematically different from the truth is biased [1]. Specifically, publication bias is defined as the tendency of investigators, reviewers, and editors to submit or accept manuscripts for publication based on the direction or strength of the study findings. This skews the public's perception of the effectiveness of the results. Therefore, any meta-analysis or literature reviews should include unpublished reports in their data instead of only including published data [2].

Numerous studies have been conducted on publication bias in general research as well as on publication bias in specific medical fields [3]. The first article with the

term “publication bias” that could be identified by searching *PubMed* was published in 1979 about the association between testicular size and abnormal karyotypes. Since then, the number of references that are potentially relevant to publication bias has considerably increased (Figure 1) [4].

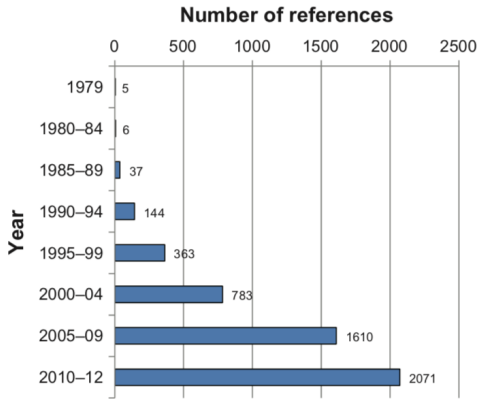


Figure 1. Number of articles relevant to publication bias identified from searching PubMed [4].

Studies with successfully proven hypothesis are represented in the literature in a greater amount than studies that “failed” to prove the hypothesis, delivering so-called negative results [5]. Prevention of publication bias by registering every trial undertaken or publishing all conducted studies is an ideal that is hard to achieve, so all stakeholders, but especially researchers, need to be conscious of disseminating negative and positive findings alike [6]. Despite the fact that several entities have taken action into combating this issue, there is still a long way to go. As an example, only 22% of the trials completed in 2009 subject to mandatory reporting by the FDA had reported their results a year after the clinical trial’s end [7].

Three main causes are highlighted which produce negative results: studies with small sample size and lacking power, no difference between groups, and more complications or adverse events in the study group [5] [8]. When the negative results are not shown, apart from an unproductive cost of time, motivation, and resources, bias in meta-analysis is introduced. This generates poor information for researchers, doctors and any readers that are interested by them [6]. Over and above scientific considerations, research participants consent to participate in research on the understanding that they are contributing to advances in treatment and scientific knowledge. The ethical duty of the researchers and editors is to honor this engagement and publish both positive and negative outcomes in an equitable manner [9].

Publication bias in the efficacy of antidepressant drugs or techniques is well-documented, as it is widely present in this important and vast field. However, there are few studies on recent publication bias in antidepressant clinical trials. That is why we formulated the following question: is there publication bias in antidepressant clinical trials from 2013-2018?

The aim of this article is to solve this research question by analyzing the statistical differences between published and unpublished antidepressant clinical trials, and to determine if its results (positive or not) had any effect on the publication of these studies. An empirical research method has been undertaken in order to answer the research question, and four distinct sections are presented (Introduction, Research Methodology, Results and Conclusion).

This analysis will lead to interesting conclusions, such as if the researchers are conditioned by the clinical trials' results when publishing them; or if publication bias has increased or decreased over the past decades.

2 Research methodology

Study Design

This study established whether publication bias is present in clinical trials involving depression that are registered on the clinicaltrials.gov website and started within 01/01/2013 to 01/01/2018. We searched all of the published clinical trials using the keywords "depression", and only selected those studies with results. We concluded that 173 trials on this website fulfilled these criteria. We defined "negative" results as non-positive, which included studies in which there was no effect, or the effect was listed as non-statistically significant.

Measurement Procedure

Solely based on the information associated with the clinical trial on the *clinicaltrials.gov* website, we determined what the study's results were (positive/non-positive), and whether a study had been published (yes/no). We methodically addressed each identified study and concluded the results from the "study results" tab, and the publication status from the "citation" section. This procedure does not consider other clinical trials involving depression that are not registered on this website, nor does it consider that some studies may not have been published because of the timeframe. Additionally, we did not perform any statistical analysis on results in which this part was absent from the website data.

Data Analysis

We analyzed the relationship between study results and publication by conducting a chi square test. This is a statistical test that establishes a statistical difference between the expected frequencies and the observed frequencies in a category.

3 Results

The objective of this study was to evaluate whether there is publication bias in clinical trials involving depression during the past five years. Out of a total sample of n = 173 clinal trials, 77 of them (44.5%) had their results published in a medical

journal, whereas the other 96 (55.5%) had not published their results on the scientific press. We determined that there is a statistically significant relationship between results and publication (p-value = 0.006). Out of the trials that are published, 29% showed non-positive results and 71% presented positive results. Out of the non-published trials, 49% showed non-positive results, whereas the other 51% were positive.

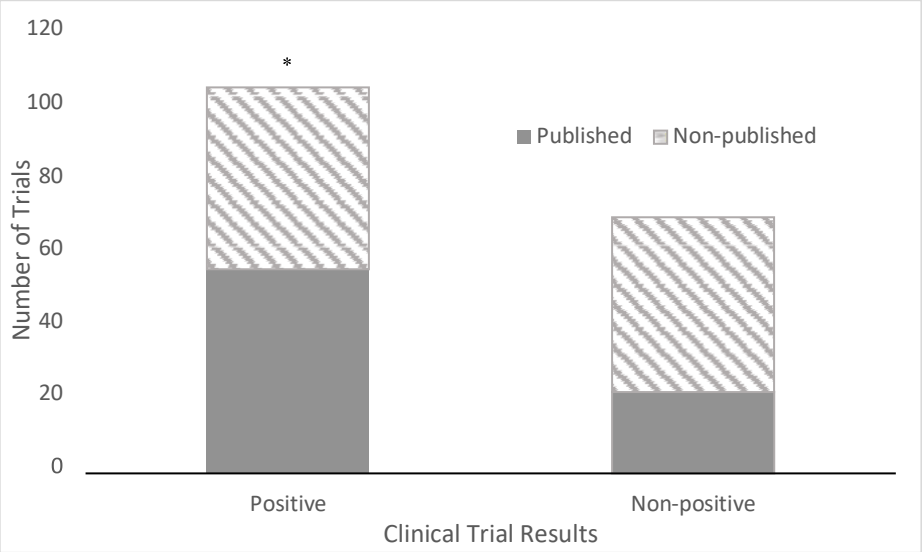


Figure 2. The number of published/non-published clinical trials with positive/non-positive results. The number of clinical trials with positive results that are published (n=49) vs. non-published (n=55) and the number of clinical trials with non-positive results that are published (n=22) vs. non-published (n=47). *, $p<0.05$ for positive vs. non-positive

4 Conclusions

The results show that there is a considerable amount of clinical trials whose results never reach the medical journals. Consequently, systematic reviews and meta-analysis on the field of depression consulting only medical press as their source will lack relevant information in order to present an accurate representation of the state-of-the-art. To avoid this bias, the authors should also consult platforms providing clinical trial results. In addition, a researcher should also ask the motives behind the non-publication of his/her results to understand the origin of the problem.

The outcome of the presented study also indicates that positive results are more likely to be published than non-positive ones. This fact could lead to the misconception of the effects of a particular therapy presented in the scientific

literature, since its beneficial effects are more likely to be published than the negative ones.

Our results also coincide with literature addressing publication bias in therapies addressed to depression [10].

The results presented in this paper conclude that there is a significant number of clinical trials in the depression field whose results are never published. Moreover, the relationship between the outcome of these clinical trials, whether positive or not, and the probability of being published in a medical journal is statistically significant. These facts endanger the credibility and representation of the state of the presented systematical reviews and meta-analysis, introducing bias in any decision based on these publications.

A limitation of our study is that our results were based on a specific website updated by the users, meaning that our conclusion could not be a veridic representation of reality if users of the platform do not properly update the state of the clinal trial. For these reasons, further research is needed in order to understand the veracity of the website.

It is also interesting to compare our results with a study that was conducted in 2000 with similar methods [11]. This study examined 74 clinical trials involving antidepressant agents that were conducted between 1987-2004. It found that ~46% of the studies were not published, and out of those ~97% had non-positive results. It is important to note that this study also identified non-positive results as results that were reported negative but came across to readers in a positive light. This number is significantly larger than our result (49%). This suggests that publication bias is indeed decreasing, which could be in part due to regulations that require all clinical trials to be registered [12].

Therefore, considering the obtained results we conclude that a considerable amount of the results for clinical trials for depression therapies conducted from 2013-2018 are not published in medical journals. This misrepresentation could result in publication bias in reviews addressing the topic. A future study could involve a more robust analysis of the published clinical trial results to examine their positiveness/non-positiveness in a statistical manner.

References

1. Otto, O., Hanninen, P., Atalay, M., Mansourian, B.P., Wojtezak, A., Mahfouz, S.M., Majewski, H., Elisabetsky, E., Etkin, N.L., Kirby, R., Downing, T.G., El Gohary, M.I.: Medical and Health Sciences - Volume VI. Eolss Publications, United Kingdom (2010)
2. Song, F., Eastwood, A., Gilbody, S., Duley, L., Sutton, A.: Chapter 21: Publication and Related Biases. In: The Advanced Handbook of Methods in Evidence Based Healthcare, SAGE Publications Ltd. pp. 371—391. United Kingdom (2001)

3. van Aert, R.C.M., Wicherts, J.M., van Assen, M.A.L.M. Publication bias examined in meta-analyses from psychology and medicine: A meta-meta-analysis. PLoS ONE. 14(4) (2019)
4. Song, F., Hooper, L., Loke, Y.K.: Publication bias: what is it? How do we measure it? How do we avoid it? Norwich Medical School. 5, 71—81 (2013)
5. Mlinarić, A., Horvat, M. Smolčić, V.S.: Dealing with the positive publication bias: Why you should really publish your negative results. Biochem Med. 27(3) (2017)
6. Thorntona, A., Lee, P.: Publication bias in meta-analysis. PlumX Metrics. 53(2), 207—216 (2000)
7. Prayle, A.P., Hurley, M.N., Smyth, A.: Compliance with mandatory reporting of clinical trial results on ClinicalTrials.gov: Cross sectional study. BMJ. 344 (2012)
8. Nair, A.S.: Publication bias - Importance of studies with negative results! Indian J Anaesth. 63(6), 505–507 (2019)
9. Joober, R., Schmitz, N., Annable, L., Boksa, P.: Publication bias: What are the challenges and can they be overcome? J Psychiatry Neurosci. 37(3): 149–152 (2012)
10. Driessen E, Hollon S, Bockting C, Cuijpers P, Turner E. Does Publication Bias Inflate the Apparent Efficacy of Psychological Treatment for Major Depressive Disorder? A Systematic Review and Meta-Analysis of US National Institutes of Health-Funded Trials. PLOS ONE. 2015;10(9):e0137864.
11. Turner, E. H., Matthews, A. M., Linardatos, E., Tell, R. A., & Rosenthal, R. (2008). Selective Publication of Antidepressant Trials and Its Influence on Apparent Efficacy. New England Journal of Medicine, 358(3), 252–260.
12. Dolgin, E. (2009). Publication bias continues despite clinical-trial registration. Nature.

Reducing Gender Bias in Natural Language Processing methods

Clothilde Breger
Ghasem Elyasi
Guillermo
Infante Mariano
Zarza

Master in Intelligent Interactive Systems
clothildeevamarie.breger01@estudiant.upf.ed
u ghasem.elyasi@gmail.com
agustinguillermo.infante01@estudiant.upf.ed
u mariano.zarza01@estudiant.upf.edu

Abstract. As in most fields today, gender bias is a problem in research and threatens the expected neutrality of research outputs. In this study we explored possible suitable methods for limiting gender bias in the research field of Natural Language Processing (NLP) by analysing the existing literature. Gender bias in NLP has two major causes: biased training corpora and algorithms' design that infer bias as a "collateral damage". We propose several methods that try to address these types of gender bias in NLP. We conclude that these methods, although very suitable by themselves, are most efficient when combined.

Keywords: Gender bias · Natural Language Processing · Debiasing methods · Algorithmic fairness.

1 Introduction

Nowadays, intelligent machines and algorithms are everywhere. One cannot deny that Machine Learning (ML) governs many aspects of the lives of people in the world's biggest economies. After all, in these societies, people create an inconmensurable amount of data everyday; for instance by using social media or credit cards. For some time now, companies and researchers have been using this data in order to draw some new knowledge about people's behaviour. This thirst for new knowledge is sometimes driven by financial gain, and other times by the desire to understand how humans interact with each other and their environment. Thus, Machine Learning techniques have been applied to extract this knowledge, most often with very high rates of success, which explains why this field is so fashionable today.

However, no technique is infallible. One of the drawbacks of Machine Learning is bias: these techniques use existing data that reflect social realities so, if the data is biased somehow (e.g. discriminating towards or underrepresenting minorities), the algorithm will produce biased output [5]. For instance, last year Amazon had

to change their recruitment algorithm after realising that it was biased against women: the use of historical data to train it meant that this program didn't see many women being successful in the company, and thus selected very few women for interviews.

One increasingly important field using more and more ML methods is Natural Language Processing (NLP): when a computer understands the meaning of a text or is able to produce a text with meaning. Then it is quite easy for ML algorithms in this field to reproduce our unconscious gender bias as they manifest in our use of language. In the 2018 IEEE/ACM 1st International Workshop on Gender Equality in Software Engineering (GE), Susan Leavy [3] gives examples of the ways in which our use of language is biased - for example an adult woman is more likely to be referred to as a girl than an adult male as a boy. In her conclusion, she states that having more women working in Artificial Intelligence, particularly considering algorithmic fairness, is essential to prevent increasingly important ML algorithms from reproducing and amplifying gender bias. This is certainly true, but it will take a while for there to be "enough" women in the field to help with the issue. For instance, [6] estimates that gender parity in Computer science publications will be reached in 2100 if current trends continue (and this is an optimistic prediction). Thus it is essential that we keep studying current methods to "debias" algorithms, all the while hoping that a new, more diverse cohort of tech researchers and workers will develop better methods in the future. Leavy indeed points out that algorithmic fairness has recently become a concern for researchers, with several papers determining the issue, and describing new methods to debias algorithms or deal with a biased training set. However, there are few papers comparing several methods - they either detail one way or list several. We thus decided to compare methods to debias algorithms in NLP proposed by four papers, in particular distinguishing those working on the corpus bias or on the algorithm design. We hope that this paper could be a step in the direction of a comprehensive list of the best methods for debiasing
algorithms, and when to use them.

2 Materials and Research Methodology

Below we present the methodology followed to develop our research, along with the materials that have supported our documentation.

2.1 Materials

Let us present our materials divided into two groups based on the way we used them in our study. On the one hand, we have a body of literature that has been the basis of our analysis, in majority composed of scientific publications from research papers to conference proceedings. On the other hand, we have all the platforms and tools used to make the literary search as accurate as possible. We have chosen four research publications to be the pillars of our analysis: [4], [8], [1], [7]. Regarding the tools that have been used to carry out this study, we have used:

– Repositories from where the literature was obtained:
  • Databases: Google scholar, The web of science, citeseer and scirus.
  • ACM, IEEE, DOI, among others
– Tools used for processing the literature:
  • JCR
  • Mendeley
  • Microsoft Excel.
– Types of literature searched:
  • Conference proceeding
  • Refereed journals
  • Refereed reviews. E.g. ACM, IEEE, DOI, among others.

2.2 Research Methodology

Our procedure for obtaining the results have been as follows. We first analysed and evaluated the general scope of gender bias in technological research, and little by little focused our research on the topic of Machine Learning, then more particularly on Natural Language Processing. Secondly, we carefully selected publications on reducing gender bias in NLP. Finally, we analysed these publications selected the most appropriate of these to review in this paper. As we could only analyse a limited amount of papers and thus only provide a first step in providing a hopefully complete analysis of debiasing methods (a daunt- ing task that will probably require several other papers), we argue that selection bias is a fatality for our paper, that we hope future researchers will correct by analysing a wider variety of papers on debiasing methods. We still wanted to limit this bias by studying papers describing different kinds of methods, espe- cially as this makes the comparison more interesting (it's not just about which method is most efficient). Therefore we chose one paper listing several methods, of which few were technical; one paper on a debiasing method based on the dataset (or corpus) given, and two papers describing more technical methods relevant to algorithm design, one having a wider array of applications than the other. See at fig:procedure a general view of our procedure.

Based on the recent definition of the procedure we have followed in our re- search, we identify our research method as within the historical methods. Of these, our methodology mostly corresponds to literature search, since examin- ing previously published studies has been one of our main tasks. This method is quite accurate when it comes to our research question and our goal of researching debiasing methods: although this method infers a selection bias due to having personally selected the literature, it is not as strong as it would be in other tech- nological fields since we are dealing with a topic that is defined and understood in the same way in all research papers (there is a consensus on the definition of bias in ML for researchers in this field). Therefore, the understanding of the environment of this topic has been much more accurate than in others, which limits the selection bias in our paper. In addition, our collection of information has been efficient, as in most cases where this method is applied.

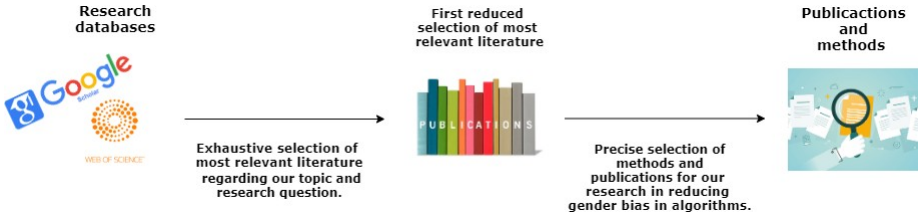


Fig. 1. General steps of our research method.

We can also approach our methodology from another perspective and offer a different view of our way of proceeding. Within observational methods, we consider that our work could be considered as a field study. Even though we have not supervised different projects, we have supervised the different methods presented in a number of publications. Thus, we have been able to obtain relevant information on how to reduce gender bias while providing an overview of the methods being researched in this field.

3 Results

The natural starting point of our research is to determine the origin of bias in Natural Language Processing algorithms. As stated before, it is clear that societal biases (like gender bias) reflected on training datasets (also called corpora) are one of the main generators of bias in the output of Machine Learning algorithms. Yet we cannot ignore problems caused by the algorithmic techniques themselves that tend to magnify gender bias. It is essential that we study both, especially because a broad method to correct bias caused by algorithm design might be more likely to solve the more latent aspects of our biases that are harder to spot. After all, to be able to debias a corpus, one must be aware not only of the population subject to bias, but also of the extent of the bias, and how it manifests itself. This why more diversity in ML and NLP workforces would help solve the issue: people who face bias can identify it better than those who don't. We will see that to debias an algorithm's design, it can suffice to know what population is discriminated against by the algorithm, which is useful as even someone who faces bias may not be able to distinguish its full extent. Note however that all of the methods we study cannot guarantee an elimination of bias altogether - there is no miracle solution.

In this section, we will analyse all papers one by one, before comparing them in our conclusion section. We start our analysis with the paper listing several methods to reduce bias, without going into detail [4]. They cite three important kinds of bias: interaction (e.g. when there are only men in the dataset), latent (e.g. there are mostly men in the dataset, so the algorithm "does not recognise" women), and selection bias (the training dataset is not representative of real population, selection was not randomised thoroughly enough). To solve these issues, the authors mention increased representation in the technical workforce,

external validity testing and auditing as possible solutions - as explained before, a greater diversity of people testing the algorithm makes it more likely that different kinds of bias will be spotted. We prefer to focus on the more technical methods listed in the paper.

- *Bias testing* : It is a simple but effective way to remove biases which already happened. By testing the model for known biases, we can try to avoid them in the future. This kind of tests should be done regularly as the model is learning.
- *Finding comprehensive data*: Having a comprehensive dataset which includes as little bias as possible is not an easy solution. For this purpose we have to pre-process the dataset and remove or correct the samples which include bias.
- *Experimenting with different datasets and metrics* : In practice, it is almost impossible to have a totally unbiased dataset due to the fact that some bias will appear after a while and we don't always have information about them at the beginning. Thus trying different datasets helps to reduce the effects of one specific bias which was included in another dataset.
- *Emotion recognition*: This method helps to resolve specific biases in facial recognition, using image recognition techniques
- *Deep learning algorithms* : Deep learning is a kind of learning from data representations. It will significantly help to reduce the effect of biases in the algorithm, as it is a very robust method.

Our subsequent papers each look at a debiasing method in detail. This one studies a method applied to the training dataset rather than the algorithms themselves. J. Zhao et al. [8] study how bias is amplified when using biased corpus in random fields based techniques. They propose a method to identify the bias, evaluate it and finally reduce it by introducing constraints so that output labels follow a desired distribution. The bias is identified by measuring the correlation between the different output variables (when one of the variables is problematic). In the case of gender bias, we want the distribution for male and female to be similar, thus the constraint is introduced to this distribution, measured by the gender ratio plus some margin. For instance, when predicting the gender associated with a certain activity (e.g. cooking) we want male and female to have the similar ratio of occurrences. One ratio might be greater than the other, but always within a margin specified by the user. The numerical results shown are positive because they manage to reduce the amplification of bias whilst protecting accuracy, but they fail to completely eliminate the bias.

The next article we review is very specific: it provides an approach to decrease gender bias in word embeddings while preserving the useful properties of the embedding [1]. The authors explain that some of the gender based information in datasets is useful; for example in the given sentence 'man is to king as women is to x' the desired value for the x is 'queen'. However for the sentence 'man is to computer programmer as woman is to x', the same models will return x equals homemaker, which is clearly a biased answer. The provided solution to overcome this situation is referred to as a debasing algorithm. It is composed of two steps.

The first one is to "identify gender subspace", that is to say determine the direction of the embedding that captures the bias. For the second step, there are two possible options: hard debiasing or soft bias correction. The former removes certain distinctions between neutral words and make them equivalent, whilst the latter reduces the differences between words but keeps more distinctions. In both cases, the goal is to find neutral words to replace the more gender-specific, biased ones. The authors evaluated the debiasing algorithms to ensure that it preserves the desirable properties of the original embedding while reducing both direct and indirect gender bias, and found that their method was approximately 13% less biased than the initial embedding.

Finally, our last paper [7] explains a general method to debias Machine Learn- ing techniques, inspired in part by the previous paper we studied. In more tech- nical terms, the process of debiasing consists of predicting some variable Y given an input vector X, while at the same time Y is debiased with respect to some protected variable Z that is a feature of X. It is worth mentioning that removing the protected variable from the input vector has been proven useless to limit bias, as shown by Hardt et al. [2]. The method explained by the authors make use of the properties of gradient-based ML techniques in order to iteratively improve the loss in the input. Depending on the type of bias, the predictor $\hat{Y}$ (i.e. the predicted values of Y) and other elements are entered as inputs of a Generative Adversarial Network (GAN) trying to predict Z, to ensure that the output stays unbiased with respect the protected variable (i.e. we cannot use $\hat{Y}$ to predict Z). The method is used for different scenarios and seems to work for all of them. Moreover, this method can be used in combination with the one proposed by Bolukbasi [1] which seems to give more flexibility to the model and enables us to use different approaches to tackle bias.

4 Conclusions

It is certainly reassuring to see that there are many methods to debias algo- rithms, even just looking at the more technical ones. None of them completely eliminate bias - especially considering the fact that we need some gender-specific information for word embeddings. It is essential to note that all methods require a more or less detailed identification of the bias we want to reduce - so we cannot repeat ourselves too much when saying that these methods must be combined with increased diversity in the field and possibly the advice of outside experts on bias. Using adversarial learning requires the least amount of knowledge on possible biases in the dataset and/or algorithm design, and it is the most gen- eral (in the sense that it can be adapted to different situations). Therefore we would recommend to use it first, and to test the algorithm for different protected variables - i.e. different characteristics the algorithm could be biased towards.

As remarked by the authors of the paper [7], to debias word embeddings this method is more flexible when combined with Bolubaski's [1], which in contrast can only be used in this context. Since we have also noticed that no method is extremely efficient in limiting bias, we would recommend associating differ-

ent kinds of methods when debiasing algorithms. As already stated, we would prescribe mixing any method with external audit or testing, but combining tech- nical methods is also possible. For instance, one could couple a general and a specific one as described above (adversarial learning and Bolubaski's method), or maybe an algorithmic method like adversarial learning with a corpus-based one as described in [8]. This is especially true as each listed method is not always usable: if one does not have full access to the dataset one cannot use the corpus- based method, or maybe we want to debias a predictor that was not obtained with a gradient-based method and so cannot use adversarial learning.This sug- gests that further research on combining these methods would be beneficial. One could study how many methods should be combined at most, or which pairs of different types of methods are more efficient.

References

1. Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *CoRR*, abs/1607.06520, 2016.
2. Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. *CoRR*, abs/1610.02413, 2016.
3. Susan Leavy. Gender Bias in Artificial Intelligence: The Need for Diversity and Gender Theory in Machine Learning. In *2018 IEEE/ACM 1st International Workshop on Gender Equality in Software Engineering (GE)*, pages 14–16, Gothenburg.
4. E. Sengupta, D. Garg, T. Choudhury, and A. Aggarwal. Techniques to elimenate human bias in machine learning. In *2018 International Conference on System Mod- eling Advancement in Research Trends (SMART)*, pages 226–230, Nov 2018.
5. Wil M.P. van der Aalst, Martin Bichler, and Armin Heinzl. Responsible Data Science. *Business and Information Systems Engineering*, 59(5):311–313, 2017.
6. Lucy Lu Wang, Gabriel Stanovsky, Luca Weihs, and Oren Etzioni. Gender trends in computer science authorship. jun 2019.
7. Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating Unwanted Biases with Adversarial Learning. *AIES 2018 - Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, 2018.
8. Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level con- straints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

An Exploration of Cross Disciplinary Approaches for Gender Debiasing in Recommender Systems

Miguel García Casado, Błażej Kotowski, Alia Morsi, Thomas Nuttall

Sound and Music Computing
miguel.garciac@upf.edu
blazej.kotowski@upf.edu
alia.morsi@upf.edu
thomas.nuttall01@estudiant.upf.edu

Abstract. Artificial Intelligence (AI) has become very intertwined in our daily lives, such that we tend to rely on decisions made by such systems in a myriad of domains. As such, not only is it important for us to evaluate the effectiveness of such algorithms beyond their typical accuracy metrics, but also to establish a common language for discussing the kinds of biases these algorithms are prone to exhibit. In this research, we explore the realm of gender bias in AI systems, specifically those relating to recommendations. Through a diverse selection of academic sources, we report some of the types of bias that are used to discuss the matter. Furthermore, we categorize, to the best of our knowledge, the different debiasing approaches encountered in our selected literature, providing the reader a decent context on which they can build their attempts to address the topic in the context of recommendation algorithms. Although our categorization for debiasing approaches is not specific to gender per se, through deductive reasoning we believe that such categories would find applicability in debiasing for all sensitive attributes in general, including gender, race or other.

Keywords: Recommendation, Recommender Systems, Gender Debiasing, Literature Search, Bias in Artificial Intelligence

1 Introduction

The pervasiveness of Artificial Intelligence (AI) in today's world puts to the forefront of public consideration the issue of algorithmic fairness. It is rather rare for algorithms to be intentionally designed to be biased, but most often, the real world data used to train such models could contain biases, incompletenesses, or discriminatory decisions [1], [2]. If models are as objective as the data on which they are trained, then it follows that inherent biases in the data will propagate to their results.

Although many types of biases exist in machine learning models, we focus our attention to those relating to gender bias only. There is a wealth of literature documenting such. Some examples include but are not limited to:

- The case of **translation** at Google [3] - where it is shown that a strong tendency towards male defaults in the translations of a large list of job titles from 12 gender neutral languages to non-gender neutral ones exists.
- Bias in **speech and face recognition systems** - where it is shown that they perform worse for women than for men [4], [5]. In [6] three commercial face recognition datasets were analyzed showing that dark-skin women are the most underrepresented group where their error rates for face recognition were up to 34.7%, compared to a maximum error of 0.8% light-skin males.
- Furthermore in **job recommendation systems** - in [7] the authors conclude, (among others) that setting one's sex to woman in Google Ad Settings, causes the advertising algorithm to show a lot less high paying jobs.
- **Music technology**. In [8] the authors argue that inclusion in Spotify playlists does affect artist's career development. [9] provides an analysis of gender distribution in Spotify playlists. They find that, although quite dependent on music genre, females occupy only from 4.5% (for "Rock This" playlist) up to 29.9% (for "Today's Top Hits" playlist). It may be that such low shares of female participation in influential playlists is due to low participation rate of females as musicians in general. However, explaining phenomenon this way does not make it more socially desirable. Technology could take a better part in reducing existing bias instead of reinforcing it.
- **Book Recommendation** - in [10] the authors attempt to understand what discriminatory bias could exist in book recommender systems. They follow an empirical approach through which they model several aspects relating to gender distributions, both in their data and in the recommender system output. They observe the distributions of recommender system output with respect to user profile data and the gender of the recommended author. They model the probability of user consumption given the author's gender, the distribution of author gender from user profile data, and the ratings of users. They also explore the ordering of names in the list of authors, and how indicative the author names are of gender. By training several different collaborative filtering models, they show that although the results differ based on the specific algorithm used for recommendation, the number of female authors read is less than the total proportion of female authors. and also that biases present in their input data were replicated in the recommender system output.

The ubiquity of this problem demonstrates a need to identify other forms of evaluation beyond the typical accuracy metrics, and to devise frameworks for detecting and addressing such biases. Proponents of such ideas include [11], who highlight that accuracy metrics are limited in that they judge the predictions of individual items rather than report results based on a set of consecutive recommendations. This gives room for many flaws to exist in the usefulness of the recommendations as an overall entity, without being detected by item-item accuracy metrics.

Recently, it was proposed to develop a new IEEE Standard on Algorithmic Bias Considerations [12] that would provide "ethical design standards that can help ensure that engineers, technologists, and the organizations they work for can provide clarity around how the algorithms they create deal with issues of bias in producing and in applying algorithms." In addition, a large number of studies aim to analyze, measure and handle algorithmic biases by proposing methods that reduce the propagation of data unfairness to the learned models. In other words, machine learning needs to be more discrimination aware. Yet, as noted by several works, it is argued that there is a lack of consensus on how to define fairness of recommendation models, and how to measure discrimination in algorithm output, and reaching such consensus would not be a trivial task [2].

However, despite ongoing efforts to demonstrate and discuss gender Bias in AI, there are very few works that contextualize the observations and proposed solutions with respect to one another, and with respect to laws concerning bias, or with frameworks for discussing bias in general. As such, based on the prior research contribution in the topic of gender bias in AI, our work aims to provide a compilation of the different types of biases that could exist in data, and aims to compare several de-biasing approaches that were applied in academia. In conducting this literature study, we follow a top-down, deductive approach. Section 2 discusses the methodology by which we gather our sources and the impact of each, section 3 concerns our results, and we conclude and highlight our limitations and future directions with section 4.

2 Research methodology

The main goal of this work is to create a framework for identifying and addressing potential unfairness in machine learning algorithms, especially that of gender bias in recommender systems. This was done by means of a literature review. However, to arrive at our results it was necessary to follow a top-down approach in reviewing the literature to better contextualize the discussions on bias in recommender systems with respect to the discussions on bias in general. Also, since we believe that some of the research done on dealing with fairness in other fields could be extrapolated to our topic of interest, it was necessary to start by reviewing the works on bias in AI in a general sense.

Our work was carried out iteratively in a couple of steps. First, to be in tune with the state of the art, we gathered a large list of references for further analysis. We started with a top down approach by collecting knowledge on general terms like bias in machine learning. Mostly google scholar and google search were utilized for our direct literature search. We identified some biases, like popularity bias, that were quite well covered in literature over almost all domains, as well as these that are less researched: like gender bias or race bias, ultimately settling on gender bias.

To move through our top-down approach and narrow down which academic contributions would be most informative for our framework, we have followed several, distinct criteria that led us to the final selection of our bibliography. First of

all, we need a good amount of references that evidence the problem (from general bias in recommender systems to specifically gender bias), and to discuss the types of bias. Then, in order to devise solutions for dealing with gender bias in other fields, we found that it is more effective to consider general approaches of debiasing and consider which ones could be applicable to our specific case. For example, since natural language processing approaches like word embeddings (defined in 3.2.2) are quite prominent in recommender system designs, and thus a number of referenced works focus on debiasing this method, then such approaches would still be relevant in the case of debiasing for gender, since it is just a specialization of the general case.

By that, we believe our approach was mostly deductive. This is because a significant portion of the bias categories and debiasing techniques encountered in the literature were not specific to particular AI applications, but were still certainly applicable to the kind of bias we are considering (gender), and to the specific kind of application (recommendation). This is because typically a proposed de-biasing approach would be applicable to a class of statistical models, and thus would apply to recommendation systems built on the same type of model.

Although it would be sensible to filter selected papers according to their impact in terms of citation counts, we did not want to restrict our reporting to only dominant works, but we also wanted to include new, original and interesting ideas, which definitely is a case of extensive literature review. So, in our choices, we left space for less impactful research efforts as well, and we were keen on maintaining some diversity. It is also worth noting that the topic of unfairness or gender disparity is a socially engaged issue, and some institutions may have a very good understanding and experience dealing with them. Therefore, some less strictly academic sources, not necessarily peer-reviewed reports like the ones from European Commission seemed to be relevant and welcome. When dealing with papers proposing possible solutions we were mostly focused on their reproducibility and the possibility of extrapolating to recommender system algorithms.

To capture diversity among our chosen sources, we included Workshop papers [1], Articles [2], [5], [15], Conference Proceedings [3], [6], [7], [16-22], Journal Papers [4], EU Technical Reports [8], [9], a Master thesis [10], an Essay [14], and a scientific Magazine Article [12]. We considered the citation counts from Google Scholar (where applicable) as a rough measure of impact, but it is worth noting that in terms of reproducibility, which is also an important metric for assessing research contributions, the most highly cited were not always the most highly reproducible. For example, the most highly cited works were [11] and [14] with 980 and 1004 citations respectively, however, they are both purely discussion papers and therefore the reproducibility metric does not even apply. On the contrary, comparing [19] and [20], with 257 and 519 citations respectively, [19] has a much higher reproducibility potential as the authors share their datasets and code, compared to [20] which gives clear description of steps but without sharing their code.

3 Results

3.1 Framework for Defining Bias/Types of Bias

3.1.1 Disparate Treatment, Disparate Impact, and Disparate Mistreatment:

In an attempt to give a framework for understanding the effect of bias in automated data driven decision making systems, [13] summarizes notions of unfairness present in the law [14], which are *disparate treatment* and *disparate impac*t. The former being present when the "system provides different outputs for groups of people with the same (or similar) values of non-sensitive attributes (or features) but different values of sensitive attributes." and the latter being present when "the decision outcomes disproportionately benefit or hurt members of certain sensitive attribute value groups". Moreover, they introduce a new notion, which they term *disparate mistreatment*, which arises as a result misclassifications due to the lack of linear separability of the training data. If misclassification rates are different for groups of people having different values of sensitive attributes, then the system could suffer from disparate mistreatment. In addition to summarizing the notions, they also formalize them in statistical terms, giving good ground for future works on addressing such notions in algorithmic models. Since there are no specific metrics to capture direct impact, the authors rely on a definition given by the U.S. Equal Employment Opportunity Commission which give a sense of direct impact quantification: the rule being that (% of subjects with a certain sensitive attribute value assigned the positive decision outcome)/ (% of subjects not having that sensitive attribute value assigned the positive decision outcome) should be no less than 0.8. This is also referred to as proportions of positive decisions, and is mentioned in [1].

3.1.2 Direct Bias and Indirect Bias

In [15] authors draw a line between direct and indirect bias. These are defined in the context of natural language processing, but still they may be useful in understanding similar phenomena in other domains. *Direct bias* is measured as a relation between two words: one gendered and one gender-neutral. Gendered words are these that may, or perhaps should be associated mostly with one gender. An example of gendered words pair would be sister and brother. Neutral words are for example mayor or architect. The stronger an association between gendered and gender-neutral words is, the stronger is the bias in such relation. Defining direct bias analogies however, does not handle the indirect bias case. *Indirect bias* may exist in data, and is expressed by close relations of words like receptionist and softball. Both of these are considered gender neutral. The fact that the word receptionist is closer to softball than to football would most likely be derived from direct bias between both gender neutral words and gendered she or woman.

3.2 Approaches for Debiasing

Here, we summarize the different debiasing techniques found in literature by classifying them based on the learning algorithm families they belong to, while noting where they are applied in the learning pipeline of the respective algorithm.

3.2.1 Debiased Gender Swap (Data Preprocessing)

Perhaps the simplest approaches to apply, these are agnostic to the specific learning algorithm since they are applied to the data itself. It is given the above title in [16], but the approach is also explored in [17]. Here, the training data is "augmented by identifying male entities and swapping them with equivalent female entities and vice-versa" hence removing correlation between gender and classification decision. This removes the incentive for the model to improve along a gender biased trajectory. Zhao et al [17] explored a similar technique in correcting gender biases in a census dataset of salary information, observing an increase in fairness at the expense of accuracy. However, the work in [18] challenges the approach that flips/swaps raw text before creating the embeddings, suggesting that such simple language modification is not as effective as repairing the word embeddings themselves. Moreover, it is circulated that these approaches should be applied with caution because the total removal of sensitive variables from the training data for debiasing could lead to forms of indirect discrimination [19]. In addition, [19] also criticizes these techniques because they could potentially lead to losses in accuracy that cannot be predicted because they treat learning algorithms like black boxes. Another critique of such methods by [1] is that even by removing the sensitive attributes among the input variables, some other variables may be correlated with the removed sensitive attribute/s. and, as a result, the classifier may capture the protected characteristics, and still induce discrimination in the decision making output.

3.2.2 Debiased Word Embeddings

A *word embedding* is a method of representing text in vector form, capturing as much useful information about them as possible in real numbers so that they can be manipulated as such (e.g. added, subtracted, multiplied). This makes them more interpretable by algorithms, and transforms the complex and unintuitive nature of structured text into a latent, lower order form of interpretable numbers. Bolukbasi et al demonstrate that gender bias related information in their system is captured across one or more of the dimensions in the latent embedding vector. [20]. With respect to debiasing word embeddings, the following approaches were encountered in literature:

a) **Hard Debiasing**. Introduced by [20] and used by [16] Hard Debiasing is an approach that first identifies the subspace across which the bias is embedded, then apply two techniques: Neutralisation and Equalization, where the

former “ensures that gender neutral words are 0 in the subspace” and the latter “perfectly equalizes sets of words outside the subspace and thereby enforces the property that any neutral word is equidistant to all words in each equality set”. In the experiments of [20], they note a 68% reduction in gender bias across there results as measured by their chosen metric (see *section 3. direct bias* for measurement and *section 4* for results impact of debiasing). This method is applied after the embedding step (referred to as a post-processing method).

b) **Soft Debiasing**. Although the Hard Debiasing approach is highly cited, many researchers challenge it and attempt to introduce softer versions [18], [17], and [16]. In [18], the authors propose reducing bias by conducting a simple linear projections for all words captured by common names. They first demonstrate biases through linear projections, and show that by attenuating this projections of some words, the bias can be slightly reduced. They propose that their linear projection technique has higher efficacy than the Hard Debiasing of [20].

3.2.3 Debiasing Convex Statistical Classifiers

[21] explores manually introducing regularization penalties in statistical classifiers such as logistic regression and in [19], the authors introduce a new measure of decision boundary unfairness, which is “the covariance between the sensitive attributes and the (signed) distance between the subjects’ feature vectors and the decision boundary of the classifier”. They claim that this novel measure enabled them to create mechanisms to train classifiers that maximize accuracy while maintaining the fairness requirement, as well as others that give flexibility to the extent of fairness they would like to introduce to the models. Moreover, they claim that disparate treatment is avoided since sensitive attribute information is not included in the decision process. They conduct their experiments with logistic regression and support vector machines.

3.2.4 Debiasing Deep Learning

Although this has not been explored deeply in the works we have reviewed, in [16], they briefly mention a technique that could have implications in deep learning based recommendation systems. If the source of learning such biases is due to the models overfitting with respect to small biased datasets with label imbalances then one approach of de-biasing could be to train the model on larger less biased datasets, and then apply transfer learning on the resultant model. This approach hypothesizes that the model would not overfit to a small biased dataset. In addition, the authors of [22] propose an adversarial learning method for increasing fairness rate of a trained network while decreasing the accuracy. They prove that even with a skewed dataset, neural network is able to learn more fair predictions. The essential aspect of this approach is tracking additional fairness metrics, apart from standard accuracy and

handle the trade-off between these two. This vision of introducing different accuracy metrics to take fairness into account goes hand in hand with the work of [11], although the work of the latter is not solely concerned with debiasing.

4 Conclusions

Having searched and analysed a wealth of literature focussing on the issue of unfairness in machine learning, we provide an aggregation of modern examples of bias in the field of machine learning and have provided an easy to follow and comprehensive offering of the most recent techniques in averting it algorithmically.

Our research has been presented with a view to demonstrate effectively the pervasiveness of the issue and make the case to the reader the need for a solution. Furthermore, we attempt to present current approaches in a way that is usable and that affords the reader the opportunity to utilise what we have deemed important for debiasing in their works, more specifically, recommender systems.

In defining the types of bias, we reported 2 classifications: one unrelated to any learning algorithm (disparate treatment, impact, and mistreatment), and another related to embeddings and how they represent data (direct vs indirect bias). With respect to debiasing approaches, we have highlighted the debiased gender swap approach (applicable at the data preprocessing phase), and we have highlighted others that are applicable when specific techniques are used. Hard and soft debiasing are examples of such techniques (applicable with word embeddings only), as are those presented for debiasing convex statistical classifiers and deep learning based approaches.

Given the findings, we believe that there is hope for devising approaches for proper debiasing, but such steps are still in infancy. We are of course aware of the limitations of our research. Taking into account the fact that we focus especially on gender bias, our proposed methods are quite general and not very directed towards solving one particular flavour of unfairness. Moreover, while we attempt to the best of our abilities to compile many approaches and compare them to one another, we concur that our comparisons are very limited due to the inability to quantitatively compare the differences in efficacy of the approaches listed. Each approach employs its own data/algorithm and the debiasing process takes on a variety of approaches and datasets. Perhaps this calls for the creation of a benchmarking standard, or set of standards where each can unify the discussion about a related task and approach for debiasing.

5 Future Work

Following the line of our methodology, the work could be quite easily extended to other flavours of bias in recommender systems. These include sensitive variables like

ethnicity, race, political views, and many more. Another possible improvement would be extending the framework of bias identification to more precise, granular categorisation. The list of methods for reducing bias could also be easily extended with additional approaches.

References

1. Žliobaitė, I. (2015). On the relation between accuracy and fairness in binary classification. ArXiv, abs/1505.05723.
2. Žliobaitė, I. (2015). A survey on measuring indirect discrimination in machine learning. ArXiv, abs/1511.00148.
3. Prates, M.O., Avelar, P.H., & Lamb, L.C. (2018). Assessing gender bias in machine translation: a case study with Google Translate. Neural Computing and Applications, 1-19.
4. Rodger, J.A., & Pendharkar, P.C. (2004). A field study of the impact of gender and user's technical experience on the performance of voice-activated medical tracking application. Int. J. Hum.-Comput. Stud., 60, 529-544.
5. Nicol, A., Casey, C., & MacFarlane, S. (2002). Children are ready for speech technology-but is the technology ready for them.
6. Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. FAT.
7. Datta, A., Tschantz, M.C., & Datta, A. (2014). Automated Experiments on Ad Privacy Settings: A Tale of Opacity, Choice, and Discrimination. ArXiv, abs/1408.6491.
8. Aguiar, L., & Waldfogel, J. (2018). Platforms, Promotion, and Product Discovery: Evidence from Spotify Playlists.
9. Aguiar, L., & Waldfogel, J. (2018). Playlisting Favorites: Is Spotify Gender-Biased?
10. Kazi, M. (2016) Exploring Potentially Discriminatory Biases In Book Recommendation.
11. McNee, S.M., Riedl, J., & Konstan, J.A. (2006). Being accurate is not enough: how accuracy metrics have hurt recommender systems. CHI Extended Abstracts.
12. A., Koene (2017). Algorithmic Bias: Addressing Growing Concerns [Leading Edge]. IEEE Technology and Society Magazine, vol. 36, no. 2, pp. 31-32.
13. Zafar, M.B., Valera, I., Gomez-Rodriguez, M., & Gummadi, K.P. (2017). Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. WWW.
14. Barocas, S., & Selbst, A.D. (2016). Big Data's Disparate Impact.
15. Chakraborty, T., Badie, G., & Rudder, B. (2016). Reducing gender bias in word embeddings.
16. Park, J.H., Shin, J., & Fung, P. (2018). Reducing Gender Bias in Abusive Language Detection. EMNLP.
17. Zhao, J., Zhou, Y., Li, Z., Wang, W., & Chang, K. (2018). Learning Gender-Neutral Word Embeddings. EMNLP.

18. Dev, S., & Phillips, J.M. (2019). Attenuating Bias in Word Vectors. ArXiv, abs/1901.07656.
19. Zafar, M.B., Valera, I., Gomez-Rodriguez, M., & Gummadi, K.P. (2015). Fairness Constraints: Mechanisms for Fair Classification. AISTATS.
20. Bolukbasi, T., Chang, K., Zou, J.Y., Saligrama, V., & Kalai, A.T. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. NIPS.
21. Kamishima, T., Akaho, S., Asoh, H., & Sakuma, J. (2012). Fairness-Aware Classifier with Prejudice Remover Regularizer. ECML/PKDD.
22. Beutel, A., Chen, J., Zhao, Z., & Chi, E.H. (2017). Data Decisions and Theoretical Implications when Adversarially Learning Fair Representations. ArXiv, abs/1707.00075.

Are young researchers trained to avoid most common biases? Guidelines for first research works.

Ainhoa M. Aguado, Elodie Medina, Marcos Mejia and Josa Prats

Master in Computational Biomedical Engineering
ainhoa.aguado@upf.edu,
{elodie.medina01,marcospaulo.mejia01,josa.prats01}@estudiant.upf.edu

Abstract. Research works are the reflection of many months of work and are very important for young researchers because they can be the beginning of a great professional career. However, sometimes due to the little work experience that some young researchers have, biases can be found (consciously or not) during the research. In order to deal with this problem, a questionnaire was made to collect the most common categories of biases and to measure the commitment of each type of bias by young researchers. From the obtained results, it was observed that young researchers are still not properly trained against biasing. This study presents a practical framework, built based on the collected results, to help to avoid biases in first research works.

Keywords: bias in research, young researcher, student.

1 Introduction

Research studies are performed for different reasons, predicting something, answering a question or improving people's well-being are just some examples of the vast amount of variety available. The problem is not the research itself but the pressure researchers undergo to publish several papers in journals during their doctoral studies due to the requests of their supervisors or studies' grant. The pressure to publish several papers and to publish them in a journal with high impact factor can lead to bias in research as stated in Uhm, C. S et al. [1]. During a study performed at University of California San Diego, it was found that 81% of young researchers in biomedical sciences had the desire to make up their results to, among other reasons, get their study published in a paper, while almost 5% of them admitted to finally made them up [2].

In the performed literature review, several references were found to have: definition of bias, specific type of bias, case examples, possible reasons for it to appear and possible solutions to avoid it. However, it is very difficult to find a good reference that collects information of all types of biases with guidelines to avoid them. The only guidelines-format reference oriented to young researchers that we have been able to find focalizes just in avoiding gender kind biases [3].

We have neither managed to find studies on the incidence of bias in the first works of researchers, which leads us to think that it is a field that has not received enough attention so far. This reinforces our idea on the need of wide meta-research work in the domain, for which our project intends to be a first step. Our hypothesis is that young researchers are not trained to avoid most common biases.

This study aims to detect which biases young researchers may be more exposed to during their first research work, and to provide them with a useful tool that details with understandable explanations and real-life examples the most common bias, as well as it proposes manners to avoid them. Bias appearance is probably closely related to the student's lack of awareness of them and can be reduced with a proper education on research.

2 Research methodology

A literature review was performed to understand all type of bias related to research in technology. From the extracted types of bias, the most common ones were selected and with them, real-life examples were built. Using real-life examples helped to build easy-to-follow techniques for students to avoid biases in their first research works.

The built examples and techniques were used to design a questionnaire, with which bias more prone to be conducted were detected. For that purpose, the questionnaire was sent to mainly Bachelor, Master and PhD students, but only those students involved or interested in research were taken into account in this study.

The questionnaire was carefully designed to collect in few questions all the main categories of bias, and to measure the knowledge of the young researchers on each type of bias, thus detecting the major weak points to address. The European standard on questionnaire development [4] was followed as far as possible.

We tried to find answers from people of different universities, nationalities and backgrounds. However, all of them were minimally related to us as the time to conduct the questionnaire was short. Also due to the short time, it was discarded to do personal interviews with current PhD students. Nevertheless, it is thought that with an anonymous questionnaire more honest and heterogeneous responses can be collected in a smaller period of time [5].

In view of the results a practical framework for future young researchers, which can be found in section 4, was developed, gathering all the scattered and difficult-to-read information from the literature, where most common types of bias and student-oriented tips are presented.

In Figure 1 the methodology described in this section is shown schematically.

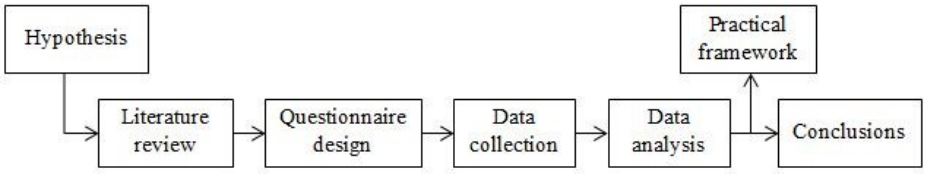


Figure 1. Methodology followed for conducting the study, starting in the formulation of the hypothesis followed by a literature review that was used to design a questionnaire.

3 Results

A total of 84 answers were collected with the questionnaire, obtaining responses of high school, bachelor, master and PhD students, as well as people currently working in a company or as researchers. From the 84 subjects questioned, just 61 were interested in research and therefore finally considered for this study, as it is aimed to serve as a guideline for researchers. In table 1, the current situation of the people interested or involved on research who answered the questionnaire is shown. It was found that 27.87% of survey respondents published a research paper, while 57.38% of them attended a research course.

Table 1 Current situation and antecedents of respondents of the survey. The quantity of people that has attended a research course is included in column *Research course* and the ones that have published a research paper are shown in column *Paper*.

Current Situation	People	Paper	Research course
High school student	5	3	2
Bachelor student	14	2	6
Master student	18	3	12
PhD student	5	3	4
Working	15	3	7
Researcher	4	3	4
Total	61	17	35

Table 2 Metric created and used to evaluate the results obtained in the questionnaire.

1. When you search for information..., in which language you look for it?		
1	0.5	0
Not in English	Just in English	Two or more languages, including English
2. Imagine you get a fancy graph that supports your hypothesis… you add a new sample and the graph does not confirm your hypothesis anymore. Would you remove that sample?		
1	0.5	0
Yes	Maybe	No
3. You create a user questionnaire... Where do you look for participants?		
1	0.5	0
In my close environment (university, friends, etc.)	All kind of people in my country...	All kind of people all over the world...
4. One patient you were analysing died... What would you do?		
1	0.5	0
Remove it from the obtained results	Include it in my analysis...	Report that this individual died during the study...
5. You are developing a platform… the results obtained for one of the samples are not accurate enough. What would you do?		
2	1	0
Hard-code it	Remove the sample	Nothing
6. When carrying out experiments that involve taking measurements with an instrument...		
1	0.5	0
I have never considered the margin error...	I know the margin error... but I do not take it into account for the results	I know the margin error... and I report it together with the results

To know how biased were the respondents of the survey the metric shown in Table 2 was created where a shortened version of the first six formulated questions and answers is shown. The scale used for this metric goes from 0 to 2, where 2 corresponds to a high bias and 0 to a non existing or very low bias. In all questions, except in question 5, a metric that goes from 0 to 1 was used. However, it was considered that hard-coding a program to obtain the desired result is more acute than removing the sample from the study. This is the reason to use another metric in this particular question.

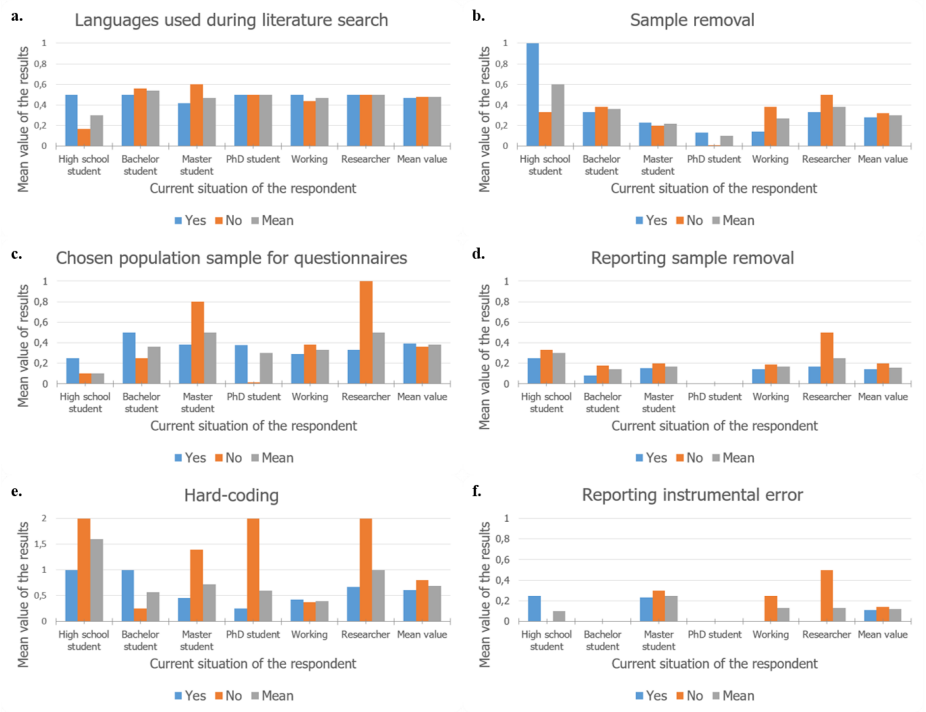


Figure 2. Mean results obtained from question 1-6, using the metric in Table 2. The blue bar corresponds to the subgroup that attended a research course, while the orange one corresponds to the ones that did not attend any research course. The grey bar represents the mean of the results obtained in each group.

In Figure 2 the mean answers (according to the metric exposed in Table 2) through the first six questions of the questionnaire are shown. Regarding to the mean in question 1 (see Figure 2a), it can be seen that the less biased group is the high school students group, in which the students that did not attend to any research course were the ones that looked for literature in two or more languages, including English. However, in the second question (see Figure 2b), related to removing a sample, the high school students' group was the most biased, even when attending to research

courses. In this particular question, it can be seen that not just the attendance to a research course helped reducing the tendency to bias, but the education received helped too.

The third question (see Figure 2c), asking to which kind of people a questionnaire would be send, was not related to the education received by the participants. Just in groups like: master students, researchers and people working, a higher biased value was obtained in the case of people not having education on research.

In question 4 (Figure 2d), most of the groups had values smaller than 0.5, and in all groups smaller mean values were acquired by participants who attended a research course. In the fifth question (Figure 2e), the most biased groups are again the ones which did not receive any education on research, while in the sixth question (Figure 2f), just in groups like bachelor and PhD students no biased answers were obtained.

Table 3 Metric created and used to evaluate the results obtained from the 7th question of the questionnaire.

7. Have you ever...	Value
Fabricated or altered data for a research work	1
Given more credibility to a source written in english over another source in another language	1
Preferred to read a paper with a positive answer to its hypothesis over another paper with a negative answer to its hypothesis	1
Given more credibility to a paper from an entity/country over another paper from another entity/country	1
None of them	0

In Table 3, the metric used to evaluate the results obtained in the 7th question of the questionnaire is exposed. The values used for this metric are 0 and 1, where 0 corresponds to a non existing bias and 1, to an existing one. All the possible responses have a metric value equal to 1 (except to last option which corresponds to “None of them”), because all the options were considered to be different types of biases (see Table 3). From the collected answers, just 12.5% of the participants of the questionnaire did not performed any of the mentioned bias. While 23.21% of them admitted to have fabricated or altered data to publish the results of their works. Related to the search of literature and the belief of the novelties presented in them, 41.07% of them admitted to gave higher credibility to a source written in English,

32.14% preferred reading references with positive answers to the stated hypothesis and 37.5% gave higher credibility to a paper from a given entity or country. Finally, 33.9% of participants admitted to have lied in questionnaires.

In Figure 3a the percentage of people that have answered the questions defined by the metric (with a scale from 0 to 2) is shown. While in Figure 3b, the answers that represent any type of bias for each question were grouped, to represent the percentage of biased answers received.

Generally, in all questions not biased answers represent at least 50% of the population, except in the first question. In this particular question, 88.52% of the participants answered that they only look for literature in English.

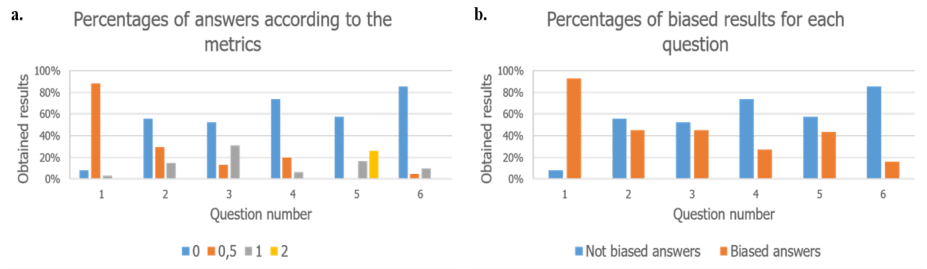


Figure 3 (a) percentage of answers received in each of the options of every question and (b) percentage of the received biased answers in each question.

Finally, in Table 4, the final mean results for each current situation is shown, differentiating the answers obtained from the ones attending to any research course or not. As it can be seen in Table 4, generally the biasing results obtained when the respondent attended to a research course are smaller, meaning that is less biased.

Moreover, depending on the current education (i.e. high school, bachelor, master, PhD) that they are receiving the results were different. PhD students had results smaller than high school students.

However, the results obtained using the different metrics (Table 2 and 3) from the researchers and workers group are higher, than the ones obtained from the PhD group. This can be due to the lack of awareness of their education level or background.

Table 4 Mean value of the results obtained in the whole questionnaire. Results are separated by the current situation of the respondent. The second column (*Yes*) shows the results of the respondent that have attended a research course, the third column (*No*) of the ones that have not attended any research course. The fourth and last column shows the mean result obtained for each current situation.

Current situation	Yes	No	Mean
High school student	4.25	4.26	4.26
Bachelor student	3.85	3.38	3.58
Master student	3.35	4.9	3.67
PhD student	2.15	3.5	2.42
Working	2.88	3.38	3.14
Researcher	3.33	7	4.25
Mean value	3.36	3.79	3.54

4 Practical framework

There exist a long list of bias, the most common and easy to find are those described below. Biases exposed are classified in different groups, each subcategory is explained in detail and accompanied with an example to exhibit the problem addressed.

4.1 Biases during information search (Availability bias)

The first category of bias that it is exposed is the availability bias category, which is one of the most difficult to detect during meta-analysis studies.

4.1.1 Language bias

One of the most common types of bias in this category is called language bias. This type of bias depends on the language of the literature searched. Song et al. [6] demonstrate that journals published in English are more likely to have greater journal

impact factors, so that previous projects written in English are usually read and found before other projects.

This type of bias can occur when the literature review is automatically searched in one language, most commonly, in English. Such behavior is explained by the importance that English has acquired throughout history in the field of research and that, at the same time, has encouraged retroactively the creation of more content in said language.

A possible solution to deal with this is not to limit ourselves to one type of language, but to look for all the papers that may have relevant information and, if needed, to use a translation tool. A systematic literature review [6] is proposed as a solution to this problematic. This literature review can be done by searching for and including grey literature, unpublished studies or data, and non-English language studies, because they can be as helpful as studies written in English.

4.1.2 Availability of sources bias

Another type of bias that also belongs to this category is called availability of sources. A clear example of this type of bias is when there is not enough and diverse information sources in order to make, a literature review.

A possible method to deal with this bias could be to search for information not only through articles available on the Internet but also through books or journals, conferences, studies of other colleagues, etc. Furthermore, in Hunter & Schmidt's book on the matter [7] it was suggested that there may be less availability bias in analyses that include primary studies examining multiple hypotheses.

4.2 Biases due to the measurement methods (Common method bias)

The next category of biases explained is called common method bias (CMB). As stated in the study of Podsakoff et al. [8], this bias happens when variations in results obtained are caused by the measurement method (e.g. an instrument) used during the research (which causes a bias, and therefore variances).

4.2.1 Errors in measurement instruments

Errors in measurement instruments are one common type of bias in this category. This bias happens when, for example, an instrument is not calibrated correctly, which can cause some input errors (e.g. a current input error for amplifiers).

In order to avoid this type of bias it is very important to be aware of the types of errors and minimum sensitivity that each used instrument can have (and the margin errors of each of them) looking at the datasheet of each instrument. This values must be taken into account when reporting the results.

4.3 Biases in experiments with human subjects

When our study includes experiments involving human subjects then we can talk about two main new categories, selection bias and intervention bias [9].

4.3.1 Selection bias

Selection bias is known as the non-correct representation of the population to be analyzed due to improper randomization during the selection of the samples. In other words, as it is often not feasible to enroll the whole population studied, it is required to select the subjects in a way that the results obtained can capture the overall characteristics of said population.

This type of error is the first and foremost to take into account whenever the collection of data is used as a medium to prove or disprove a hypothesis (empirical research) and it should encompass most of the attention at early stages of our investigation. Moreover, a flawed sampling selection could lead to undesired misled conclusions, and since the magnitude of this bias' impact and the direction of its effect is unpredictable not much can be done once it occurs [10,11].

Therefore, the best way to deal with this bias is to minimize it during the early phases (recruiting of individuals) and have special care while retaining the sample population whenever a follow up is required [10,12].

To illustrate this error, we will use this example retrieved from [13]:

> "*A hypothetical case-control study was conducted to determine whether lower socioeconomic status (the exposure) is associated with a higher risk of cervical cancer (the outcome). The "cases" consisted of 250 women with cervical cancer who were referred to Massachusetts General Hospital for treatment for cervical cancer. They were referred from all over the state. The cases were asked a series of questions relating to socioeconomic status (household income, employment, education, etc.). The investigators identified control subjects by going from door-to-door in the community around MGH from 9:00 AM to 5:00 PM. Many residents are not home, but they persist and eventually enroll enough controls.*"

The problem becomes evident when we compare the obtention of data in both groups. We can identify that the method used for the control data may have tended to select inadvertently individuals of a specific socioeconomic status as women staying at home at that hours were more likely to be unemployed.

All these circumstances lead into a sample that is not representative of the whole population. To avoid it we must try to include subjects from different origins. For instance, in the case of questionnaires it is relatively easy to reach people from all

over the world through the internet, and even to find more people that meet a specific profile of our interest (e.g. automotive enthusiasts) in thematic communities.

Volunteer/non-volunteer differences are more difficult to address, but it can be useful to design the experiment in the least possible invasive way in order to maximize the probability of recruiting participants who would have not accepted if they had felt uncomfortable.

4.3.2 Intervention bias

The intervention bias category includes phenomena related to the way in which experiments are developed. Intervention bias can happen when the subject is aware of the intent of the study and tends to give more favourable responses or perform better, or when differences exist in the setup or explanations of the experiment between different subjects, among others.

We can also find the so called contamination bias, when our study is based in the comparison of two groups involved in two different experiments, if one or both groups are exposed to features of the other experiment, or just if the subjects are aware of the differences the study is researching, and this influences their behaviour.

To avoid this kind of bias it is always important to assign subjects and experiments randomly to groups, and to be very systematic in the experiment development repeating the same setup and explanations for all of the subjects. It is also very recommended not to reveal the specific aim or hypothesis of the study until the end of the experiment.

4.4 Biases during data analysis (Confirmation bias)

Confirmation bias is another common category of bias. This bias describes the tendency to manipulate data to obtain a result that supports the hypothesis. R. S. Nickerson et al [14] divided this category into two, depending on the awareness of the researcher on the manipulation of the data. While the first one implies deliberately building a case to justify a conclusion, the second one is done unconsciously.

This type of bias is quite dangerous, as often researcher give less importance to the evidence that refute their theory. Mahoney M. J [15] stated that one of the reasons by which it could appear is the pressure researchers face when their work needs to be published as soon as possible.

4.4.1 Fabrication and falsification bias

When facing results that do not support the hypothesis of the study, usually these are ignored or are given less importance than the ones that support it. A possible

method to avoid this kind of bias is to include all the results of the experiments and if a sample is removed, report it. Furthermore, the validity of a hypothesis is statistical, so it is not a problem to have a few negative results.

The fabrication and falsification of evidence is easy to avoid, as the manipulation is an action done consciously. It happens when evidence is changed to maintain coherence with the desired results.

The technique proposed to avoid this type of bias is to have always revised the acquired evidence, at least twice, and have it revised by colleagues not related to the study. Having a person disconnected from the research and that does not know which is the hypothesis of the work will help to see evidence objectively avoiding like this the unconscious action of ignoring undesired evidence.

4.5 Biases in the publication of research work (Publication bias)

Finally, publication bias is an additional type of bias that we should bear in mind. Publication bias occurs when articles or research reports are less published in research communities or the media for reasons like the language in which they are written, the presence of a controversial central topic, methodology or conclusions, a negative answer to the hypothesis raised, or even reasons beyond the research as the age, the race or the gender of the authors [16]. This type of bias is the only one that does not depend on the researcher, but on the people in charge of the journals, conferences and other media, so we can do little to really prevent it from happening. However, it can be useful to know about it, both to take into account that it can affect us in the form of availability bias when we search for information, and to try to avoid some of its causes (e.g., the language of the article) in our own work.

5 Conclusions

For this study a questionnaire on research practices was designed and more than 80 answers from young people close to a research environment were collected. Participants included high school, bachelor, master and PhD students, as well as people currently working in a company or in research. However, only the participants interested in research were considered for the study, summing a total of 61 subjects.

A metric system was established to estimate the biasing scores, and from the analysis of the given results we can infer that young students are not currently trained enough to approach their first research works with guarantees to generally avoid common biases. A difference was detected between participants who had assisted some kind of course on research practices and participants who had not. However, the high biasing scores got for all categories in both groups show that further education is required on this direction.

Interesting findings from this study hint that the most generalized type of bias is language bias, from the category of information searching, since most of the participants look for information only in English, losing the chance of finding useful data from sources written in other languages. On the other hand, the less committed type is the common method bias associated to instrumental error in measurements.

A relevant fact detected is a reduction on the biasing tendence for PhD students, which can be associated to their longer experience in research, although the data collected from this group is limited since only 5 participants were PhD students. In addition, the level of education of participants currently working in a company is unknown. The results obtained from high school students are doubtful as 60% of them claim to have a paper published and 40% of them attended a research course. Furthermore, the questionnaire was not sent to any high school students, so it is probable that a misunderstanding during the selection of this option happened.

Our project aims to be a first step approach towards bias education and avoiding. Further work would include testing the understandability and acceptation of the proposed practical framework in real environments of young research. After this, it would be interesting to validate its effectiveness by comparing the research works of students who have been trained with the framework and those who have not.

References

1. Uhm, C.-S., Waterman, M.S.: What Is Research Misconducts? Publication Ethics Is as Important as Research Integrity. Applied Microscopy. 46, 67--70 (2016)
2. Fanelli, D.: How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. PLoS ONE. 4(5) (2009)
3. Jelić, N.: Bias in conducting research: Guidelines for young researchers regarding gender differences. Journal of European Psychology Students. (2013)
4. Brancato, G., Macchia, S., Murgia, M., Signore, M., Simeoni, G., Blanke, K., Hoffmeyer-Zlotnik, J.: Handbook of recommended practices for questionnaire development and testing in the European statistical system. European Statistical System. (2006)
5. Johnson, B., Turner, L. A.: Data collection strategies in mixed methods research. Handbook of mixed methods in social and behavioral research, 297--319. SAGE Publications, Thousand Oaks (2003)
6. Song, F., Parekh, S., Hooper, L., Loke, YK., Ryder, J., Sutton, AJ., Hing, C., Kwok, C. S., Pang, C., Harvey, I.: Dissemination and publication of research findings: An updated review of related biases. Health Technol Assess. 14(8) (2010)
7. Hunter, J. E., Schmidt, F. L.: Availability and source bias in meta-analysis. Methods of meta-analysis, 493--510. SAGE Publications, Thousand Oaks (2004)
8. Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., Podsakoff, N. P.: Common method biases in behavioral research: A critical review of the literature and recommended remedies. Journal of Applied Psychology. 88(5), 879--903 (2003)
9. Krishna, R., Maithreyi, R., Surapaneni, K M.: Research Bias: A Review For Medical Students. Journal of Clinical and Diagnostic Research. 4(2), 2320--2324 (2010)
10. Odgaard-Jensen, Jan, et al.: Randomisation to Protect against Selection Bias in Healthcare Trials. Cochrane Database of Systematic Reviews. 4(1), 1--11 (2011)

11. Nour, S., Plourde, G.: Pharmacoepidemiology in the Prevention of Adverse Drug Reactions. Academic Press (2019)
12. Berger, V. W., Christophi C. A.: Randomization Technique, Allocation Concealment, Masking, And Susceptibility Of Trials To Selection Bias. Journal of Modern Applied Statistical Methods. 2(1), 80--86 (2003)
13. Boston University School of Public Health, Selection Bias in Case-Control Studies. http://sphweb.bumc.bu.edu/otlt/MPH-Modules/EP/EP713_Bias/EP713_Bias-TOC.html
14. Nickerson, R. S.: Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. Review of General Psychology. 2(2), 175--220 (1998)
15. Mahoney, M. J.: Publication prejudices: An experimental study of confirmatory bias in the peer review system. Cognitive therapy and research. 1(2), 161--175 (1977)
16. Song, F., Hooper, L., Loke, Y.: Publication bias: What is it? How do we measure it? How do we avoid it?. Open Access Journal of Clinical Trials. 5(1), 71--81 (2013)

A. Annex

The questionnaire designed for this study is the one that follows:

A practical framework to avoid biases in your first research work

This questionnaire on research practices is completely anonymous, please answer with the utmost sincerity, otherwise the recollected information would not have any validity.. Thank you very much!

1. **Current Situation**
   - High school student
   - Bachelor student
   - Master student
   - PhD student
   - Research assistant
   - Researcher
   - Employee / Freelance worker
   - Another choice

2. **Are you involved or interested on research?**
   - Yes
   - No

3. **Have you published any research paper?**
   - Yes
   - No
   - Another choice

4. **When you search for information (e.g. papers) for your research, in which language you look for it?**
   - English (English is my native language)
   - English (English is NOT my native language)
   - Another choice

5. **Imagine you get a fancy graph that supports your hypothesis. However, you try it with a new sample and the graph does not confirm your hypothesis anymore. Would you remove that sample?**
   - Yes
   - No
   - Maybe

6. **You create a user questionnaire for your research work. Where do you look for participants?**

- In my close environment (university, friends, etc.)
- All kind of people in my country, fulfilling the required profile if it exists
- All kind of people all over the world, fulfilling the required profile if it exists
- Another choice

7. **One patient you were analysing died before you finished your study. What would you do?**
   - Remove it from the obtained results
   - Include it in my analysis, even if I don't have all the required follow ups
   - Report that this individual died during the study and remove it from the obtained results

8. **You are developing a platform that works well for almost all patients. However, the results obtained for one of them are not accurate enough. What would you do?**
   - Remove the sample
   - As I know how I can obtain the desired result, hard-code it for this sample
   - Nothing

9. **When carrying out experiments that involve taking measurements with an instrument…**
   - I know the margin error of the instrument and I report it together with the results
   - I know the margin error of the instrument but I do not take it into account for the results
   - I have never considered the margin error of the instrument
   - Another choice

10. **Have you ever...**
    - Lied when answering a questionnaire
    - Fabricated or altered data for a research work
    - Given more credibility to a source written in english over another source in another language
    - Preferred to read a paper with a positive answer to its hypothesis over another paper with a negative answer to its hypothesis
    - Given more credibility to a paper from an entity/country over another paper from another entity/country
    - None of them

Effectiveness of Methods for Evaluating Technology

Henry Hasti[1], Kohki Mametani[2], Ana Gabriela Pandrea[3], Yiqun Liu[4]

Master in Sound and Music Computing
[1]henry.hasti01@estudiant.upf.edu,
[2]kohki.mametanii01@estudiant.upf.edu,
[3]anagabriela.pandrea01@estudiant.upf.edu,
[4]yiqun.liu01@estudiant.upf.edu

Abstract. Evaluating the quality of a software is an important process for both software developers and users. In this paper, we compare the states of research behind two common approaches for software evaluation: human-based and artificial intelligence (AI)-based. We propose a weighting matrix using four criteria: human usability, machine usability, verification, and validation. The results show the two research tracks have different strengths in different areas. Therefore, we conclude that the most proper research scheme is the tailored approach combining both human and AI testing.

Keywords: Software, Evaluation, Artificial intelligence

1 Introduction

Software plays an essential role in virtually everything humans do: virtually all engineering projects rely on software in some form, as do government and bureaucratic functions. Even trivial and social interactions rely on software that meets a particular goal and works as it is stated. Therefore, the quality of a software is an important issue that we should care about, including the general usability, sustainability, and maintainability. To evaluate a piece software is a tricky balance between hard objectivity and the very subjective (but very valid) individual user experience. The main research goal of this study is to answer the question of how users can verify and trust the quality of a software program within the scope of how technology evaluation is currently researched. We analyze the reasons to validate software as well as the methods for doing so within the context of existing technology evaluation research.

Evaluation of software is a well-studied problem. In the past, researchers concentrated on the reasons to do it and the methods of doing it. Nowadays, the state-of-art technology combines these considerations with artificial intelligence (AI). These approaches, frequently termed diagnostic classification [1], involve training the neural networks on data that human reviewers of software have created so that the networks can automatically detect errors in software and verify it. These methods are much more efficient than human programmers/testers, and can also be applied to new software more readily, but they face the downside of not being very understandable: it is difficult for a human to understand the approach of the network, which in turn can complicate the verification process [2].

In this paper, we attempt to delve into these approaches and analyse the research behind them. In the following section we propose a weighting matrix to evaluate strengths and weaknesses

between the two research approaches. In the third section we complete this matrix, and we conclude in the last section.

2 Research methodology

We propose performing dual literature reviews to determine leading technology/software evaluation methods that rely on either human-based testing or AI-based testing. We will then investigate the research fields behind each and compare their strengths and weaknesses. We hope to research potential shortcomings in the existing research fields of software evaluation and AI evaluation. Given our findings in previous work, we will use the following documents to inform our survey of human evaluation methods:

- [3]
- [4]

While these documents do not exhaustively encompass the work in the field of software evaluation research, we believe that they provide a highly representative sample of evaluation methods, particularly given the scope constraints of this paper. Other sources involved in our initial literature review generally overlap in terms of the evaluations prescribed. Using the combined methods presented by the two papers (the first paper focused on computational accuracy and the second on usability and user experiences), we arrive at the following researcher-preferred methods to evaluate software:

- Testing: Give the software a test case with a known result and compare the actual output to the expected
- Wrapping: Build the software into an outer software module that checks whether the inner module behaves as expected for every input
- Informal proofs: Confirm that the math behind the software works appropriately
- Numerical error estimation: Check the maximum likely error due to propagation of errors from measurement, numerical instabilities, and rounding
- Survey: Ask users about their experiences after using the software
- User testing: Users are monitored (where they look, what gestures they make, e.g.)
- Heuristic evaluation: Experts evaluate the software usability on certain heuristics

We will use the following documents to inform our analysis of AI evaluation. As before, they are not exhaustive but rather representative.

- [2]
- [5]

The papers describe research work on AI methods to review both the style and output of code.

Our methodology is to compare the research methods proposed for human reviewers to those proposed for AI. We will acknowledge the problems that may arise through following the human approaches, and compare them to problems from the AI by completing the weighting matrix

below, with the goal to more quantitatively determine the present advantages of each method and research path over the other.

Human usability is defined as how usable the software should be for humans, and essentially coincides with the evaluations proposed in Paz et. al. Machine usability is defined as how usable the software is with other programs. Verification should check that results are mathematically correct. Validation should check that the program implements the correct functions. Importance is based on our subjective interpretation following our literature review, but the values can easily be changed to meet the needs of future research. Likelihood of failure measures how likely either the human or AI reviewer is to incorrectly evaluate the issue. Catastrophic-ness measures how bad an extreme error could be from either method. The total is calculated as

$$\text{Importance} * (\text{Likelihood} + \text{Catastrophic-ness}) \quad (1)$$

Adding the totals gives the total negative score for the given method, and is our metric for comparing the human and AI research methods.

Table 1 Weighting matrix. We propose to calculate the effectiveness of human and AI methods by completing this table

Issue to be tested	Importance	Likelihood of failure with given method	Catastrophic-ness of failure with given method	Total
Human usability	**7**			
Machine usability	**5**			
Verification	**10**			
Validation	**9**			

3 Results

Results of our research show that most of the technical papers lack a trustworthy level of certainty that their proposed software is valid in terms of both approaching the problem right and solving it right. While some authors give written logical arguments and motivations, they do not emphasize the use of complete test suites, either manually written or AI based.

In regard to the effectiveness of evaluation methods, we have discovered that both human and computer tests provide useful results depending, of course, on the nature of the evaluated piece of

software and its scopes. While a human evaluator is more suitable for a human-computer interaction (HCI) system, a computer based - either specifically programmed or learnt with AI - system would be more suitable for complex mathematical and computational tasks.

Another important result was the good practice of peer reviewing, which gives us some hope in what concerns the quality of innovative software that is being proposed in research. While the expertise of these peers is still questionable, findings show that with the help of AI that learns and generalizes from these reviews, evaluation of software in research could be an accessible tool for everyone. One of these tools is DeepCodeReviewer (DCR), as described in [5]. Its architecture is in the picture below.

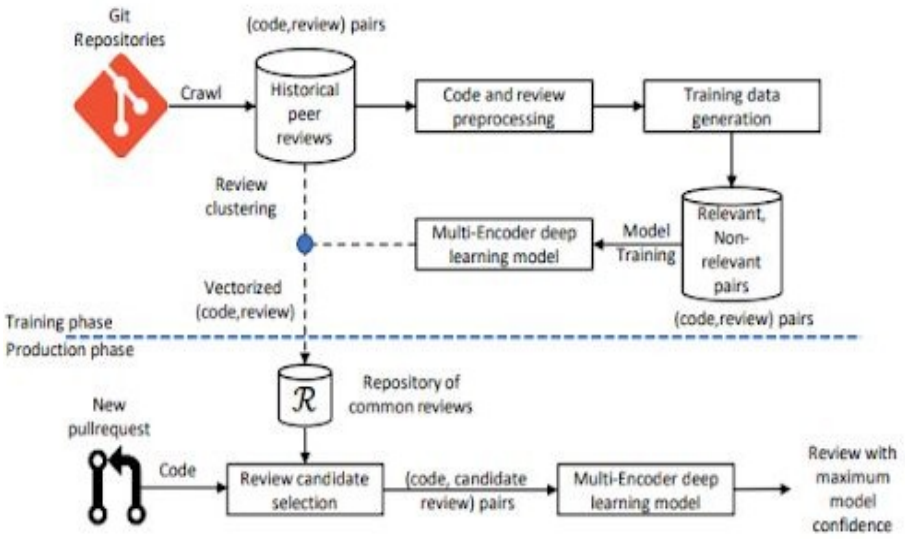


Figure 1: Overall architecture of DeepCodeReviewer With two main phases: training phase processes historical peer reviews and trains a deep learning code review model; production phase makes use of trained model to perform code analysis and apply relevant review. [5]

Under all the described circumstances, we completed Table 1 as a general guideline for both human and AI evaluators. The results show how an AI evaluator is more stable with a lower likelihood of failure. From the total scores we can note that the AI evaluator exceeds the human one in all the tasks apart from Human Usability, which is understandable since humans should be able to evaluate this aspect better and more natural.

Table 2: Effectiveness of human evaluation. The total is 470

Issue to be tested	Importance	Likelihood of failure with given method	Catastrophic-ness of failure with given method	Total
Human usability	**7**	**3**	**5**	**56**
Machine usability	**8**	**8**	**7**	**120**
Verification	**10**	**7**	**8**	**150**
Validation	**9**	**8**	**8**	**144**

Table 3: Effectiveness of AI evaluation. The total is 350

Issue to be tested	Importance	Likelihood of failure with given method	Catastrophic-ness of failure with given method	Total
Human usability	**7**	**5**	**5**	**70**
Machine usability	**8**	**4**	**6**	**80**
Verification	**10**	**3**	**8**	**110**
Validation	**9**	**2**	**8**	**90**

We also predict that, by combining the two independent methods, the evaluation results could improve significantly. This means that the most appropriate testing environment for a proposed piece of software will combine human and AI test design on a case by case basis, depending on the nature and scope of the program in question.

4 Conclusions

In this work, we compared the software evaluation methods proposed for human reviewers to those proposed for AI by evaluating the quality of each research method against 4 criteria: human usability, machine usability, verification, and validation. As for the usability, our investigation revealed that the usability of software evaluation methods is highly dependent on the nature of software to evaluate. The specific trends we discovered are that the evaluation by human gives the highest usability in the context of HCI systems while the computer-based evaluation is the most suitable for mathematical and computational system. In regard to verification and validation, we found that most of the software evaluation methods lack a way to verify that their proposed method is rightfully approaching the question they want to solve. The main reason behind the low certainty is the insufficient awareness of the use of complete test suits among the research community. The significance of our research is that this work provides researchers of software evaluation and software developers with an analytical viewpoint to understand what evaluation method is the most suitable for their system, and how researchers of both human- and AI-based can more effectively carry out their work.

References

[1] Y Adi, E Kermany, Y Belinkov, O Lavi, Y Goldberg, "Analysis of sentence embedding models using prediction tasks in natural language processing", *IBM Journal of Research and Development*, vol. 61, no. 4, pp. 3-1, 2017.

[2] L. Liu, Z Jiang, "Research on software reliability evaluation technology based on BP neural network," in 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS)*, 26-29 June 2016, Okayama, Japan* [Online]. Available: IEEE Xplore, https://ieeexplore.ieee.org/abstract/document/7550924. [Accessed: 9 Nov. 2019].

[3] United States. US Department of Transportation, Federal Highway Administration, *Software Reliability: A Preliminary Handbook*. McLean: Department of Transportation;2004. [Online]. Available: https://www.fhwa.dot.gov/publications/research/safety/04080/04080.pdf. [Accessed: Nov. 9, 2019].

[4] F. Paz and J.A. Pow-Sang "A Systematic Mapping Review of Usability Evaluation Methods for Software DevelopmentProcess" *International Journal of Software Engineering and Its Applications,* vol. 10, pp. 165-178, 2016.

[5] Microsoft Corp., "Intelligent Code Reviews Using Deep Learning," *KDD'18 Deep Learning Day*, 2018. [Online]. Available: https://www.kdd.org/kdd2018/files/deep-learning-day/ DLDay18_paper_40.pdf. [Accessed: Nov. 9, 2019].

Research and Validation Methodologies for Music Technologies

Miguel Pérez, Jorge Bustos, Roberto Pérez, David Bedoya

Sound and Music Technologies Master Program
{miguel.perez, jorge.bustos, roberto.perez, josedavid.bedoya}01@estudiant.upf.edu

Abstract. Applying good research methodologies is essential within science. In computer science specifically, the past 20 years researchers have emphasized on following good practices of research methodologies within this field. Research groups have to follow carefully these methodologies for doing proper research. In this paper, we analyze the main research and validation methodologies in computer science and how these methodologies are being applied in the music technology (MT) field and in particular in the Music Technology Group (MTG) of Pompeu Fabra University. Overall, the performed analysis provides some insights about the main validation methodologies the music technology field and the MTG are applying.

Keywords: Validation Methodology; Research Methodology; Music Technology; Music Technology Group; Pompeu Fabra

1 Introduction

During the 21st century, research in MT and the evaluation of different methods and techniques have evolved to become one of the widest research technology fields in our days.

MT is an enormous field of study. Considering all the different features and researches made in this topic, the top five most researched topics in MT could be the following [1]: (i) Music Generation/Modeling, (ii) Sound Generation/Modeling, (iii) Music Performance Analysis/Synthesis, (iv) Music Interfaces and Music and (v) Audio Understanding/Retrieval.

There also exists a good variety of methods for technology research and evaluation. Zelkowitz, M. V. et al. [2] developed a list of 12 different experimental approaches for validating technology, as shown in table 1, and classify them into three different categories: observational, historical and controlled methods. But, which of these 12 approaches are the ones used in MT?

Research methodologies have been evaluated mostly for a specific technology or field, for instance, Gulati, S. et al [15] evaluated methodologies for melodic similarity in audio recordings of Indian art music, Urbano, J. et al [16] focused the evaluation in the specific field of Music Information Retrieval (MIR).

However, this paper centers its attention in the evaluation methods in the MT field in general to extract insights about which are the most used and verified research and

evaluation methods and compared them with the ones used in the MTG research group from Pompeu Fabra University. To do so, first we describe the different research and evaluation methods proposed in [2] and its application of music and sound technology. Then a variety of MT and MTG research papers is analyzed and the results are shown. Finally, by comparing both results we provide some observations and conclusions.

2 Research methodology

When researchers tackle some technology in order to get some results or products from their studies, they can follow four different approaches toward this experimentation [2]:

- Scientific method: It is based on the verification or refutation of a given hypothesis.
- Engineering method: Given a hypothesis, a solution for it is developed or tested. Then, based on the results, this solution is improved until no improvement in possible
- Empirical method: Data is collected from a statistical point of view in order to validate a hypothesis, which may not be described by a formal model or theory
- Analytical method: Results are concluded from a formal theory and compared with empirical observations.

In order to achieve our research goals, we follow an empirical research methodology, where we collect different papers related to MT from various sources, trying to derive a conclusion about which are the research and validation used in those papers. The main features of the different validation methods by Zelkowitz [2] are also explained in **Table 1**.

We also consider another validation method exposed in [3]: surveys. A survey focuses on obtaining the same kinds of data from a large group of people (or events) in a standardized and systematic way.

In order to see and make a conclusion about the research and evaluation methodology in the MT research environment, we analyzed 11 different papers about the trending topics in MT research nowadays. 7 (63.64%) of these papers proceed from various sources and 4 (36.36%) of them were written and developed in the MTG in Pompeu Fabra University.

Table 1. Summary of validation methods [2]

Validation method	Description	Weakness	Strength
Project monitoring	Collection of development data	No specific goals	Provides baseline for future; Inexpensive
Case study	Monitor project in depth	Poor controls for later replication	Can constrain one factor at low cost
Assertion	Ad hoc validation	Insufficient validation	Basis for future experiments
Field study	Monitor multiple projects	Treatments differ across projects	Inexpensive form of replication
Literature search	Examine previously published studies	Selection bias; Treatments differ	Large available database; Inexpensive
Legacy	Examine data from completed projects	Cannot constrain factors; Data limited	Combine multiple studies; Inexpensive
Lessons learned	Examine qualitative data from completed projects	No quantitative data; Cannot constrain factors	Determine trends; Inexpensive
Static analysis	Examine structure of developed product	Not related to development method	Can be automated; Applies to tools
Replicated	Develop multiple versions of product	Very expensive; "Hawthorne" effect	Can control factors for all treatments
Synthetic	Replicate one factor in laboratory setting	Scaling up; Interactions among multiple factors	Can control individual factors; Costs moderate
Dynamic analysis	Execute developed product for performance	Not related to development method	Can be automated; Applies to tools
Simulation	Execute product with artificial data	Data may not represent reality; Not related to development method	Can be automated; Applies to tools; Evaluate in safe environment

3 Results

Table 2 shows a review of the first seven papers analyzed in this study. They include different areas regarding MT, such as Deep and Machine Learning, or MIR.

Watching and understanding the analysis of these seven papers, we can see that all of them uses an engineering methodology in order to execute the research work. We can also see a trend in mostly all of them when talking about the evaluation method. All of them uses an evaluation method based in the execution of the designed system or research in the paper (either dynamic analysis or simulation). This execution evaluation was compared with legacy data from related projects in 5 out of 7 papers analyzed.

Table 2. Analyzed papers not belonging to the MTG group.

Paper	Area	Source	Date	Citations	Research Method	Objective Evaluation	Subjective Evaluation
[4]	Deep Learning	ISMIR	jun-19	0	Engineering Method	Legacy Data - Dynamic Analysis	N/A
[5]	Machine Learning	IEEE	may-08	17	Engineering Method	Legacy Data - Dynamic Analysis	N/A
[6]	Machine Learning	Journal of New Music Research	sept-14	12	Engineering Method	Dynamic Analysis	N/A
[7]	MIR	IEEE	dic-15	27	Engineering Method	Legacy Data - Dynamic Analysis	Survey
[8]	MIR	ACM	oct-06	122	Engineering Method	Dynamic Analysis	N/A
[9]	MIR	Journal of New Music Research	may-16	15	Engineering Method	Legacy Data - Simulation	N/A
[10]	MIR	ISSCR	nov-18	1	Engineering Method	Legacy Data - Simulation	N/A

However, if we look to the papers developed in the MTG, we can notice some differences with the other seven. Although the research method used in the MTG is also an engineering methodology and the evaluation method is also based on the execution of the program, the subjective evaluation differs from the other investigations: The survey plays a fundamental role in the evaluation methodology in order to perceptually validate the truth or functionality of the system or research being developed in the different papers.

Table 3. Analyzed papers from MTG

Paper	Area	Source	Date	Citations	Research Method	Objective Evaluation	Subjective Evaluation
[11]	Source Separation	ICASSP	mar-19	0	Engineering Method	Legacy Data- Dynamic Analysis	Survey
[12]	Deep Learning	ICASSP	may-19	4	Engineering Method	Simulation	Survey
[13]	MIR	IEEE	dic-16	0	Engineering Method	Simulation	Survey
[14]	Deep Learning	IEEE	ene-18	49	Engineering Method	Simulation	Survey

4 Conclusions

After analyzing 11 papers in various fields of MT, we can see one thing in common: they use an engineering methodology. This point makes sense since the field of MT itself is one in which existing technologies and knowledge are exploited to find solutions to related problems. We can also note that in MT research outside the MTG of the Pompeu Fabra University, surveys are rare, limiting the evaluation of the proposed solution or method to benchmarks, artificially generated data or other data used in similar technologies. However, in the MTG, surveys are much more common, and it seems that they will remain so because old and modern papers continue to maintain them. It seems that there is not direct relationship between the method used and the number of mentions, and that the success of the paper among other investigations is determined by how popular the topic is (such as deep learning, as we can see in [14][12], which are very recent papers but with a moderate number of citations) or by establishing a fundamental model such as in [8]. It is important to state that these conclusions do not generalize well since the amount of papers is not very representative given the numerous literatures existing on this field.

For future work a bigger quantity of papers in the field of MT must be considered. In addition, it might also be interesting to compare the evaluation of MT in other departments that investigate these issues throughout the world.

References

1. Serra, X. (2005). Towards a roadmap for the research in Music Technology. *International Computer Music Conference, ICMC 2005*.
2. Zelkowitz, M. V., & Wallace, D. R. (1998). Experimental models for validating technology. *Computer*, *31*(5), 23–31.
3. Oates, B. J. (2006). Researching Information Systems and Computing. Inorganic Chemistry.
4. Choi, K., & Cho, K. (2019). Deep Unsupervised Drum Transcription.
5. Wang, J., Wu, Q., Deng, H., & Yan, Q. (2008). Real-time speech/music classification with a hierarchical oblique decision tree. ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 2033–2036.
6. Hedges, T., Roy, P., & Pachet, F. (2014). Predicting the Composer and Style of Jazz Chord Progressions. Journal of New Music Research, 43(3), 276–290.
7. Davies, M. E. P., Hamel, P., Yoshii, K., & Goto, M. (2014). AutoMashUpper: Automatic creation of multi-song music mashups. IEEE/ACM Transactions on Audio Speech and Language Processing, 22(12), 1726–1737.
8. Harte, C., Sandler, M., & Gasser, M. (2006). Detecting harmonic change in musical audio. Proceedings of the ACM International Multimedia Conference and Exhibition, 21–26.
9. Bernardes, G., Cocharro, D., Caetano, M., Guedes, C., & Davies, M. E. P. (2016). A multi-level tonal interval space for modelling pitch relatedness and musical consonance. Journal of New Music Research, 45(4), 281–294.
10. Bernardes, G., Davies, M. E. P., & Guedes, C. (2018). A Hierarchical Harmonic Mixing Method. In M. Aramaki, D. M. E. P., R. Kronland-Martinet, & S. Ystad (Eds.), Music Technology with Swing (pp. 151–170). Cham: Springer International Publishing.

11. Chandna, P., Blaauw, M., Bonada, J., & Gómez, E. (2019). A Vocoder Based Method For Singing Voice Extraction.
12. Blaauw, M., Bonada, J., & Daido, R. (2019). Data Efficient Voice Cloning for Neural Singing Synthesis. 6840–6844.
13. Parekh, S., Font, F., & Serra, X. (2016). Improving Audio Retrieval through Loudness Profile Categorization. 565–568.
14. D. Rethage, J. Pons and X. Serra, "A Wavenet for Speech Denoising," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, AB, 2018, pp. 5069-5073.
15. Gulati, S., Serrà, J., & Serra, X. (2015). An evaluation of methodologies for melodic similarity in audio recordings of Indian art music. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 678–682.
16. Urbano, J., Schedl, M., & Serra, X. (2013). Evaluation in Music Information Retrieval. *Journal of Intelligent Information Systems*, *41*(3), 345–369.

A Survey of Music Information Retrieval Evaluation Practices since 2013

Georges Naimeh, Şiyar Vurucu, Jorge Marcos Fernández

Sound and Music Computing Masters, Universitat Pompeu Fabra, Barcelona
{georges.naimeh01, siyarramazan.vurucu01, jorge.marcos01}@estudiant.upf.edu

Abstract. Evaluation practices in Music Information Retrieval have been a focus of the discipline for a long time now. In 2013, a series of suggestions for improving evaluation of research in MIR was published. Here, we examine the adoption of some of those suggestions in research published in ISMIR from 2014 to 2019. We find that data sharing seems to have increased, and find this correlates with the increasing use of deep learning techniques in MIR. Our findings are not exhaustive or conclusive, but serve as a starting point for further research on the topic.

Keywords: MIR, evaluation, research practices, open research, experimental methods.

1 Introduction

Music information retrieval (MIR) is a relatively young and highly multidisciplinary research field, started around two decades ago [1]. It is concerned with the "extraction and inference of meaningful features from music [...], indexing of music using these features, and the development of different search and retrieval schemes" [1]. Technological developments starting in the late 1990s and early 2000s, such as greater computing power of personal computers, and greater and more widespread availability of music players and streaming services (e.g. Spotify), have contributed to its success as a field in its own right.

However, given its relatively recent creation, the field has had to slowly conform to established research practices. This has required a significant amount of research on research practices in MIR (i.e. not research on MIR tasks, but how this research is conducted). Already in 2004, Downie observed the necessity for standardised collections of data and agreed upon MIR task definition so researchers could scientifically compare their systems [2]. But he points to even earlier acknowledgements of this need in MIR research:

"The MIR community has long recognized the need for a more rigorous and comprehensive evaluation paradigm. A formal resolution expressing this need was passed on 16 October 2001 by the attendees of the Second International Symposium on Music Information Retrieval (ISMIR 2001). (See music-ir.org/mirbib2/resolution for the list of signatories.)" [2]

In 2006, Flexer [3] stated the importance of statistical procedures in MIR research evaluation. Looking at the 2004 proceedings from the leading conference in the field, ISMIR (International Society for Music Information Retrieval), he found that "only 6 papers [out of 53 to which statistical methods were applicable] reported mean performances plus standard deviations to give an idea of the variability of results", and "only 2 papers [of those 53] employed a statistical test to prove the significance of their results."

More recently, Urbano et al. [4] covered the challenges faced by the MIR community regarding evaluation procedures in great detail and, based on an exhaustive literature review and their own experience in the field, compiled a list of proposed changes to the way research is conducted in the field. This will be the starting point of the research presented here. According to [5] "Music is listened to, performed and created by people. It is therefore essential to consider the user as central to the creation of user scenarios, hence to the development of technologies." However, since evaluating user experience directly is expensive, impractical and poses difficulties for reproducibility, it is the system response that is actually evaluated in MIR research, and therefore a correlation between system response and user experience is assumed [4]. This assumption underlying the MIR evaluation process motivates the contribution of Urbano et al. [4] on the grounds of its implications on research validity, reliability, and efficiency.

The main aim of this paper is to produce an assessment of current MIR research practice, especially as it concerns evaluation procedures, to help guide further research on and implementation of MIR evaluation practices that are conducive to reliable and reproducible research. Specifically, we survey the uptake of six measures proposed by Urbano et al. [4] and discuss the implications of the trends shown by the data.

2 Research Methodology

This paper uses a historical approach to assess to what degree evaluation practices are improving in MIR according to the measures proposed by Urbano et al. [4]. Since the ISMIR conferences are considered the main publication in the MIR community, we randomly sampled 10 papers from each year's proceedings since 2013 (when [4] was published) and searched for keywords related to each issue being considered.

The issues under consideration were chosen on the basis of how difficult they were to count as a binary factor; i.e. yes/no issues. Furthermore, as [6] recently evidenced, and supported amongst others by Urbano et al.'s [4] main claim, there is a data problem in MIR research: "The ability of researchers to verify each other's work is an integral step in the scientific process; without it, consensus on new findings is difficult to reach". Thus, reproducibility is diminished [7], which threatens reliability. For these reasons, issues related to sharing were deemed important to assess in our survey: code availability, data availability, and results sharing. Additionally, statistical significance was chosen because of its effects on reliability and validity [3][4]. Another point highlighted by [4] is that of standardisation in general, and software standardisation in particular. Given the introduction in 2014 of the mir_eval software package to calculate standard MIR evaluation metrics [8], and the ease of consideration in surveying ISMIR papers (i.e. does the paper use mir_eval?) we added it to our list of issues.

Lastly, we also added the use of deep learning into consideration. This is simply a control measure, given that a greater availability of data could be due to the recent surge in deep learning research ("deep learning revolution" [9]), an inherently data-driven discipline, and not to a greater awareness of data usage and its implication on reproducibility in the research community.

The six issues in MIR evaluation from those identified by Urbano et al. [4] taken into account in this paper are described as below. Only binary values are possible (yes/no, 1/0).

- Code availability [CA]
    - 1: the paper shared the code on a public repository or offered it on demand.
    - 0: the paper did not mention its public availability in any way.
- Data availability [DA]
    - 1: the paper used a publicly available dataset for either training (if applicable) or testing/evaluation, or created one and mentioned its public availability online.
    - 0: the paper used private dataset(s) for any research stage.
- Result sharing [RS]
    - 1: the paper provides an extended table/figure with proper statistical analysis of results and/or provides unedited results as a separate link to publicly available repository

    - 0: the paper only provides average scores on evaluation metrics and does not share a link to published raw results
- Use of the mir_eval package [ME]
    - 1: the mir_eval package was used and cited
    - 0: the mir_eval package was not used in evaluation of the MIR systems
- Use of statistical significance methods [SS]
    - 1: measurements such as p-values, t-tests, or similar are employed in result analysis.
    - 0: no such measurements are carried out on results.
- Use of deep learning [DL]
    - 1: the paper's main contribution entails the use of deep learning techniques.
    - 0: the paper does not employ deep learning in any considerable way.

The raw data can be found in the Appendix.

3 Results

In Figure 1, we see the increase in number of deep learning oriented papers. At ISMIR 2019, more than half of the papers are using deep learning practices in various tasks in MIR.

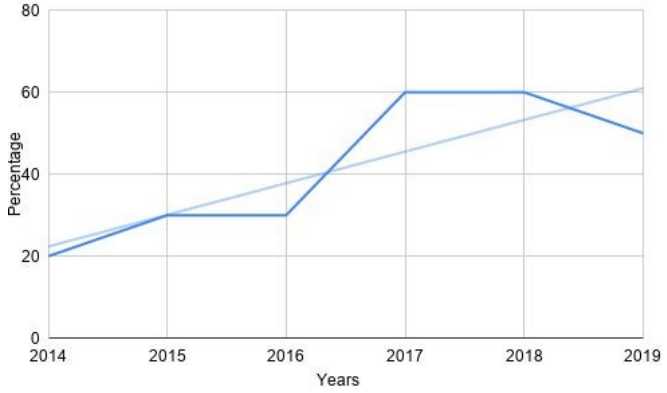


Figure 1. Percentage of papers using deep learning practices in ISMIR across the years

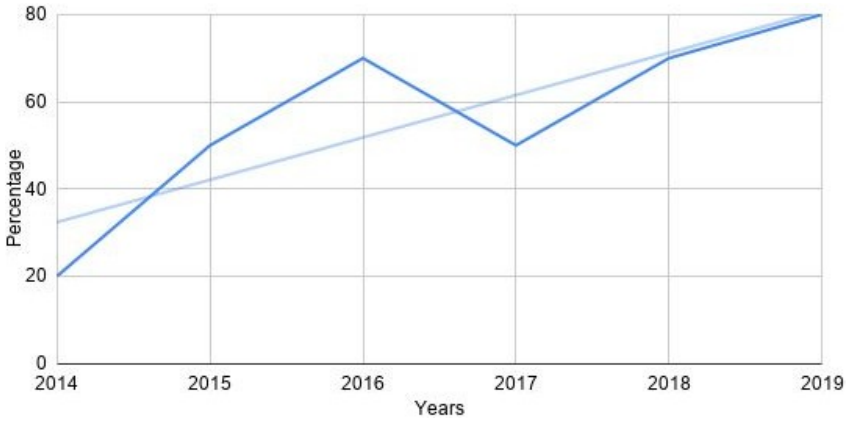


Figure 2. Percentage of papers sharing their evaluation/training data (DA) in ISMIR across the years

A drastic increase of data sharing throughout years can be observed in figure 2. There are mainly two factors for this increase. First, community becomes more aware of the importance of reproducibility. Second, data sharing percentage increases simply because the percentage of studies using data driven methodologies are increased (Figure 1). There is a positive correlation between deep learning papers and data availability (Table 1). But since 1) The number of papers sharing their data increased more than the increase in number of deep learning papers (Figure 1, 2) and 2) Data availability in Non-DL papers is actually higher DL papers by a small margin (Figure 6), we conclude that awareness of reproducibility problem has increased in MIR field.

Table 1: Correlation between issues of ISMIR papers

DL	SS	CA	DA	ME	RS	
1	-0.55079	0.33183	0.55543	0.23702	-0.34696	DL
	1	0	0	0	-0.37796	SS
		1	-0.30237	0.2	-0.73192	CA
			1	0.52915	-0.27664	DA
				1	-0.58554	ME
					1	RS

Papers using deep learning also share their code more than others. First reason for this result is, it is more crucial for DL papers to share their code, because the methods they use regard to deep learning for a specific MIR task is also subject to research and evaluation. Whereas authors of Non-DL papers usually explain the algorithm mathematically and not bother with sharing their code.

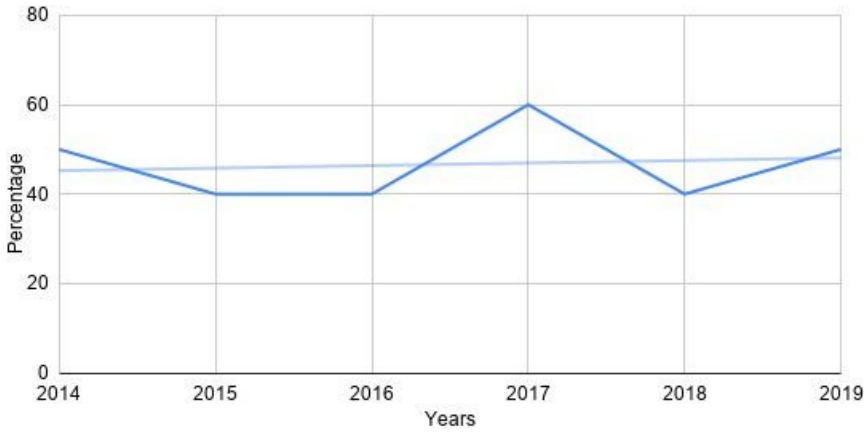


Figure 3. Percentage of papers sharing their software implementations (CA) in ISMIR across the years

Code availability does not increase across the years (Figure 3). One of the factors is that the number of DL papers authored by companies is in increase and they do not share their software implementations, which hinders the ratio of code availability.

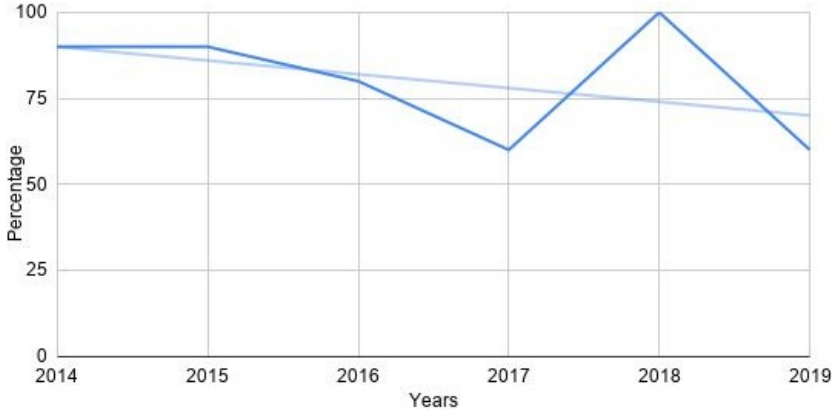


Figure 4. Percentage of papers sharing their results (RS) in ISMIR

Decrease in result sharing is partly due to newly defined tasks with no established baselines to compare.

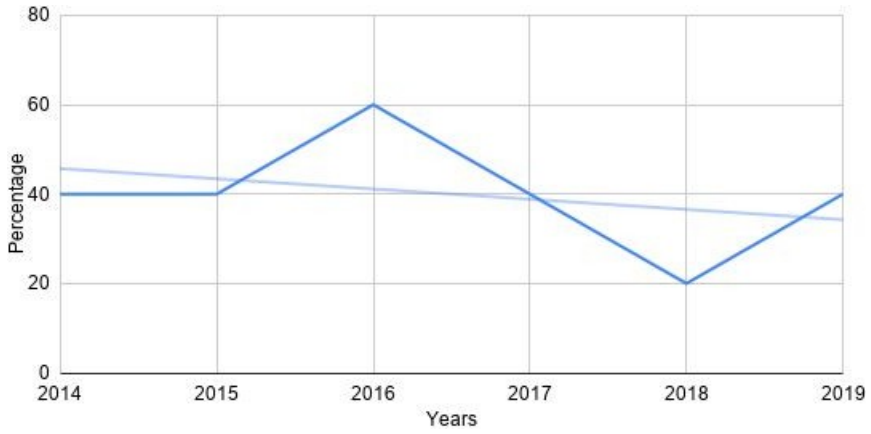


Figure 5. Percentage of papers providing statistical significance (SS) for their results in ISMIR across the years

The ratio of papers using statistical significance measures to validate their results decreased (Figure 5). Correlation between deep learning papers and use of statistical significance is -0.55, (Table 1) , and number of DL papers using SS is nearly half of the Non-DL papers (Figure 6). Deep learning models have a great number of parameters to be tuned [10]. SS measures are important to validate the effects of those parameters, which is ignored by the authors according to our results.

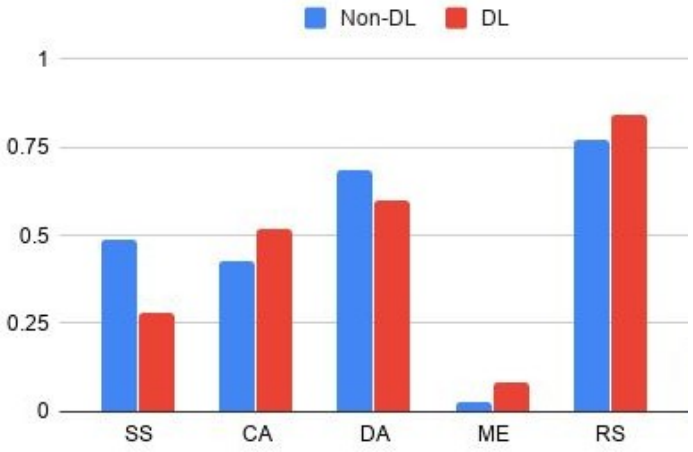


Figure 6. Comparison of DL and Non-DL papers across other issues

4 Conclusions

In this paper we have outlined some of the issues surrounding evaluation of Music Information Retrieval Systems. We tried to measure how healthy is the current evaluation methods by looking at statistical significance measures, code availability, data availability, result sharing rate and use of standard software package mir_eval. We observed that DL papers are increasing the code and data availability but lacks using statistical measures.

In future work, we would like to increase the number of papers analysed. With some simple scripts, the analysis task could be automated, and looking at the entirety of ISMIR publications from each year would be feasible. This would yield more reliable results. We would also encourage analysing the adoption of some of the other suggestions in [4], especially those to do with dataset creation and sharing (e.g. the use of multimodal data).

References

1. Schedl, M., Gómez, E. & Urbano, J. *Music Information Retrieval: Recent Developments and Applications*. *Foundations and Trends® in Information Retrieval* **8**, (2014).
2. Downie, J. S. The scientific evaluation of music information retrieval systems: Foundations and future. *Computer Music Journal* **28**, 12–23 (2004).
3. Flexer, A. Statistical evaluation of music information retrieval experiments. *Journal of New Music Research* **35**, 113–120 (2006).
4. Urbano, J., Schedl, M. & Serra, X. Evaluation in music information retrieval. *Journal of Intelligent Information Systems* **41**, 345–369 (2013).
5. Serra, X., Magas, M., Benetos, E., Chudy, M., Dixon, S., Flexer, A., Gómez, E., Gouyon, F., Herrera, P., Jordà, S., Paytuvi, O., Peeters, G., Schlüter, J., Vinet, H., Widmer, G. *Roadmap for Music Information Research*. Editor: Peeters, G. Creative Commons BY-NC-ND 3.0 license, ISBN: 978-2-9540351-1-6 (2013).
6. Chen, W. *et al.* Data Usage in MIR : History & Future Recommendations. *ISMIR 2019* 25–32 (2019).
7. Vandewalle, P., Kovacević, J. & Vetterli, M. Reproducible research in signal processing: What, why, and how. *IEEE Signal Processing Magazine* **26**, 37–47 (2009).
8. Raffel, C. *et al.* mir_eval: A transparent implementation of common MIR metrics. in *Proceedings of the 15th International Society for Music Information Retrieval Conference, ISMIR 2014* 367–372 (2014).
9. Sturm, B. L. Revisiting priorities: Improving MIR evaluation practices. *Proceedings of the 17th International Society for Music Information Retrieval Conference, ISMIR 2016* 488–494 (2016)
10. Bergstra, J. S., Bardenet, R., Bengio, Y., & Kégl, B. (2011). Algorithms for hyper-parameter optimization. In Advances in neural information processing systems (pp. 2546-2554).

Appendix: Analysed Papers

Year	ID	DL Paper	Statistical Significance	Code Available	Data Available	Mir_eval usage	Results Shared	DOI
2014	1	0	0	0	1	0	1	10.5281/zenodo.1417925
2014	2	0	0	0	1	0	1	10.5281/zenodo.1416150
2014	3	1	0	0	1	0	1	10.5281/zenodo.1416084
2014	4	0	0	0	0	0	1	10.5281/zenodo.1417081
2014	5	0	1	1	1	0	1	10.5281/zenodo.1415566
2014	6	1	0	1	0	0	1	10.5281/zenodo.1416944
2014	7	0	0	1	1	0	1	10.5281/zenodo.1417595
2014	8	0	1	1	1	0	0	10.5281/zenodo.1418013
2014	9	0	1	NA	0	NA	1	10.5281/zenodo.1417993
2014	10	0	1	1	1	0	1	10.5281/zenodo.1417091
2015	1	0	0	1	1	0	1	10.5281/zenodo.1416824
2015	2	1	1	0	1	0	1	10.5281/zenodo.1416968
2015	3	0	1	1	0	0	1	10.5281/zenodo.1415582
2015	4	0	0	1	1	0	1	10.5281/zenodo.1415728
2015	5	0	1	0	0	0	1	10.5281/zenodo.1417751
2015	6	0	0	0	0	0	1	10.5281/zenodo.1417044
2015	7	0	0	0	1	0	0	10.5281/zenodo.1417103
2015	8	1	0	1	0	0	1	10.5281/zenodo.1415806
2015	9	1	0	0	0	0	1	10.5281/zenodo.1417531
2015	10	0	1	0	1	0	1	10.5281/zenodo.1414996
2016	1	1	1	1	1	0	1	10.5281/zenodo.1418305
2016	2	0	1	0	0	0	1	10.5281/zenodo.1417073
2016	3	1	1	1	1	0	1	10.5281/zenodo.1417819

2016	4	0	0	0	1	0	1	10.5281/zenodo.1417825
2016	5	NA	NA	NA	NA	NA	NA	10.5281/zenodo.1417801
2016	6	0	0	0	0	0	0	10.5281/zenodo.1414924
2016	7	0	1	1	1	0	1	10.5281/zenodo.1417659
2016	8	0	1	0	1	0	1	10.5281/zenodo.1414968
2016	9	0	0	1	1	0	1	10.5281/zenodo.1414724
2016	10	1	1	0	1	0	1	10.5281/zenodo.1417739
2017	1	1	0	1	1	0	0	10.5281/zenodo.1417427
2017	2	1	0	1	0	0	1	10.5281/zenodo.1417937
2017	3	0	1	0	1	0	1	10.5281/zenodo.1417193
2017	4	1	0	1	1	0	1	10.5281/zenodo.1418015
2017	5	0	0	0	0	0	0	10.5281/zenodo.1417567
2017	6	1	1	1	1	0	1	10.5281/zenodo.1416370
2017	7	0	1	0	0	0	0	10.5281/zenodo.1416188
2017	8	1	0	0	0	0	0	10.5281/zenodo.1417737
2017	9	1	0	1	0	0	1	10.5281/zenodo.1415990
2017	10	0	1	1	1	0	1	10.5281/zenodo.1417000
2018	1	1	0	0	1	0	1	10.5281/zenodo.1492337
2018	2	1	0	0	0	0	1	10.5281/zenodo.1492347
2018	3	0	0	1	1	0	1	10.5281/zenodo.1492357
2018	4	0	0	1	1	0	1	10.5281/zenodo.1492367
2018	5	1	0	1	1	0	1	10.5281/zenodo.1492377
2018	6	1	0	0	1	0	1	10.5281/zenodo.1492389
2018	7	1	0	0	0	0	1	10.5281/zenodo.1492399
2018	8	1	0	0	0	0	1	10.5281/zenodo.1492411
2018	9	0	1	1	1	0	1	10.5281/zenodo.1492419

2018	10	0	1	0	1	0	1	10.5281/zenodo.1492429
2019	1	1	0	1	1	0	1	10.5281/zenodo.3527741
2019	2	1	1	0	1	1	0	10.5281/zenodo.3527898
2019	3	1	0	1	1	0	1	10.5281/zenodo.3527866
2019	4	0	1	0	1	0	1	10.5281/zenodo.3527856
2019	5	0	0	1	1	1	1	10.5281/zenodo.3527870
2019	6	1	1	1	1	1	1	10.5281/zenodo.3527766
2019	7	0	0	0	0	0	0	10.5281/zenodo.3527790
2019	8	0	NA	1	1	NA	NA	10.5281/zenodo.3527756
2019	9	1	0	0	0	0	0	10.5281/zenodo.3527840
2019	10	0	1	0	1	0	1	10.5281/zenodo.3527816

Public engagement in Personalized medicine: A comparative study

Lieke Ceton
Mar Galofré
Paula Lampreave
María Prado

Computational Biomedical Engineering Master´

liekejohanna.ceton01@estudiant.upf.edu mar.galofre01@estudiant.upf.edu
paula.lampreave01@estudiant.upf.edu maria.prado01@estudiant.upf.edu

Abstract. Public engagement is an important component in the implementation and understanding of new technologies in medicine. It is vital in personalized medicine, a technique that requires a high level of patient information and communication. A survey was performed on 63 European participants to analyze their perspective on the subject and compare it to an earlier study on the American public. The results show that both groups are predominantly positive and interested in the development. However, a significant group feels ill-informed and questions are raised in a wide range of topics. Investing in public engagement will be well-received and could help answer their questions. This could aid both the implementation and the development of personalized medicine.

Keywords: Personalized Medicine, Public Engagement, Innovation in Medicine.

1 Introduction

Public engagement describes the two-way process of sharing new developments in higher education and research with the big public. This interaction results in a society that is more knowledgeable and supportive of new technological advances. Feedback of the public can also contribute to more relevant and desirable innovation, causing new techniques to be implemented more easily. Good communication between researchers and the public has important mutual benefits [1].

This study focuses on public engagement in personalized medicine, an area in the medical field where the communication between the researcher and the patient is vitally important. Personalized or precision medicine refers to treatments that are tailored to the individual patient, ideally providing the right treatment at the right time, offering greater accuracy and efficiency than traditional treatments. It is an approach that requires a lot of patient data like detailed family history and genetic information [2]–[4].

Therefore, the technique requires a high level of public trust. In 2018, the Personalized Medicine Coalition together with GenomeWeb performed a study on the perspectives of Americans on personalized medicine. The participant rate of this study was 1001, the sample was randomly drawn from a large national Survey Sample Panel of U.S. adults [5].

The main fields they studied were [5]:

- Their knowledge of personalized medicine.
- Their reaction to a description of personalized medicine
- Their belief about whether insurance companies should cover personalized tests and treatments.
- Their concerns to apply for personalized medicine.

They conclude, amongst others, that the public is generally positive about the matter. However, some part of the population is hesitant in sharing this information for fear of it being stolen or used in a decrement way for the individual. [2][3]. At this moment, public misunderstandings and fears about the new technology might possibly hold back its clinical implementation[2]–[4].

The aim of this paper is to do a comparative study, analyzing the perspective of the European public in personalized medicine and comparing the results of the previous study realized outside of Europe. Its results can be used to design fitting public engagement strategies to enhance the level of engagement.

2 Research methodology

A representative survey has been conducted with 63 European participants in the Autumn of 2019. Google forms [6], from Google Docs, was used to develop the survey. The general structural was taken from the American survey but the questions were adapted for the European public. The information was then analyzed using Microsoft Excel [7].

The data collected from the survey was used to extract some conclusions about personalized medicine in Europe and to compare it with the perspectives on personalized medicine in America. Once that the comparison was performed, some ideas for further improvement of public engagement in personalized medicine have been extracted.

The survey was shared via different social networks. At the end, 63 subjects fill out the complete questionnaire, which was composed of 3 parts with different questions each one[8].

The first part consists of general information about the participants (age, nationality, sex, the higher degree of education and medicine background). This part has been used to know whether the survey sampling made the survey truly valuable, reliable and representative [9].

The second part was composed of questions related to what they already knew about personalized medicine in order to measure awareness of and opinions about the field and its stated benefits. Participants were asked after their confidence level, giving a number on their knowledge in the topic.

The last part consists of some questions after learning what personalized medicine is. This part is useful to study public engagement in personalized medicine. It has been analyzed whether after reviewing information about itself and its benefits, the participants were more interested in the field and wanted to learn more about it.

At the end of the study, all the data obtained through the surveys was collected and analyzed to extract results and to compare these results with the ones obtained at the American survey [10].

3 Results

As mentioned before, the survey used in this study was answered by 63 participants. The group consisted of 44 females and 19 males with a mean age of 22 (SD 2). The vast majority have a European nationality, mainly Spanish (57%). Four answers came from non-European nationalities, all who are living in Europe.

The educational background of each subject was also recorded in the survey. In order to acquire this information, the participants were asked about their highest degree of education. The most common answer was Bachelor's degree (48) followed by Master Degree (8). However, only 17.5% of the group had degrees related to medicine.

After completing the data regarding describing the population, the next question focused on the main topic of this paper, personalized medicine. Firstly, participants were asked if they knew what personalized medicine is. They were given a 1-5 scale, one meaning they did not have any information on the topic and 5 they were confident in their knowledge. The mean value for this answer is 2.59 (SD 1.28). The mode, as seen in **Figure 1** is 1, meaning that the majority of the subjects don't have information about personalized medicine.

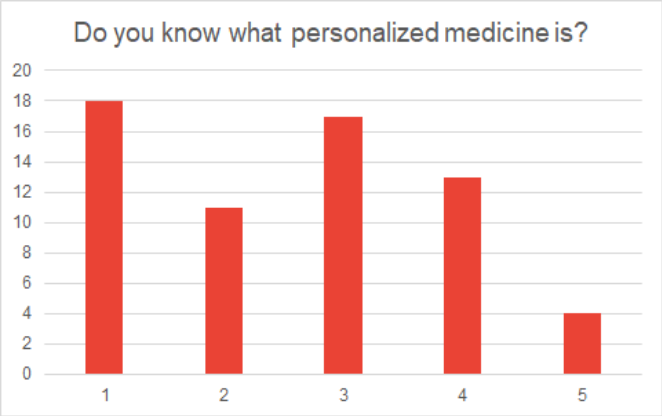


Figure 1. Graph representing the knowledge of participants in the topic of personalized medicine

Following this, participants were asked to choose three concepts out of a list that described personalized medicine better for them. The three top concepts chosen were 'Better for the patient', 'Safe' and 'Time saving'. **Figure 2** summarizes all the answers obtained.

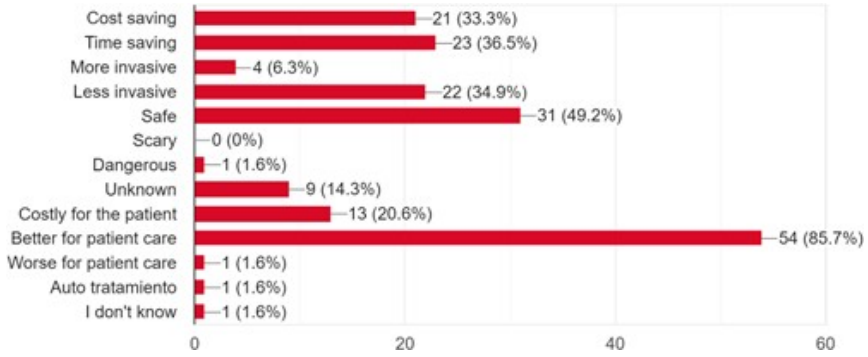


Figure 2. Graph summarizing the chosen concepts regarding assumptions about personalized medicine

This concludes the first part of the survey, where no information was given to the subjects. Before answering the next set of questions, a brief explanation of personalized medicine and some of its applications was given. Then, the participants were asked if their confidence level on the topic had risen. Same as before, they were asked their knowledge about this area of medicine, and the mean answer given is 3.86 with a standard deviation of 0.76. Also, the mode, as seen in **Figure 3**, has risen to 4, showing an improvement compared to the first time this question was asked.

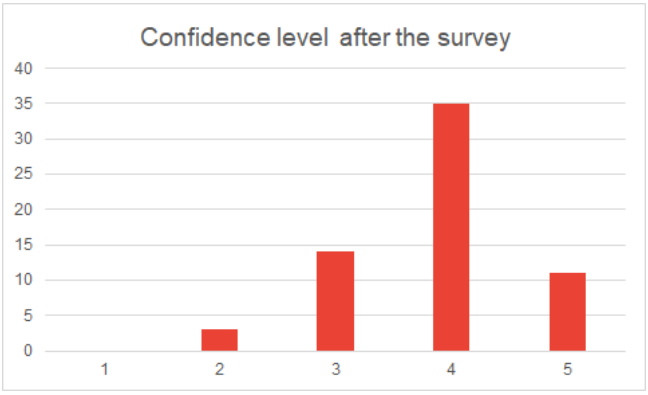


Figure 3. Graph representing the knowledge of participants in the topic of personalized medicine after being giving a small piece of information

In the same line, participants had add if they feel that their opinion about the topic change after being given a small piece of information, only 39.69% of the subjects confirmed it change, commenting that they understood better what personalized medicine is, or how it works. Few participants added that they were unaware of the genetic component.

Focusing more in the public engagement area of this piece of work. Only 44% of the participants thought they were given enough information about this area of medicine before taking this survey. However, 93.65% of the participants added that they would follow a personalized treatment or procedure if it was offered by a healthcare professional.

To final question given was open answer. Participants had to give a few ideas of concepts or knowledge they would like to have about personalized medicine. Answer focuses on a variety of topics:

- Costs for the patient and the healthcare system
- Safety and risk of the genetic analysis and the actual procedure or treatment
- More information about the illness that can be treated with personalized medicine

4 Discussion

The American survey draws four key findings from their results. The first finding states that the public is not familiar with the subject. The participants of this survey show a confidence level of 2.6/5 on beforehand, implying that they neither feel very confident nor very insecure about the subject beforehand. After a short description this level increases by more than a full point up to 3.86/5, showing similarly that the concept was not understood fully by the public.

The results show that the participants have a positive and interested attitude towards the subject. The vast majority feels that it would be better for the patient and almost all would undergo a treatment themselves. In the written comments, key words like useful, efficient and revolutionary are found. This is similar to the American study, which reports that most of the public is excited about the field.

The little reported concerns are spread over a number of topics including data storage and surgical danger. Contrary to the American study, they are not focused mainly on the insurance costs. This can be explained by the differences in the healthcare system of America, where there is not access to universal care whereas in Europe universal healthcare system isseen across the countries . It is reported that one third of the European public feels the treatment will be cost saving. More than concerned, the public is interested in the subject, although half feel ill-informed and want to know more.

The fact that this survey was answered only by young, highly educated participants (more than 80% has a university degree) could have an influence on the way personalized medicine is received. Researching a representative sample of the European public will give a better overview of the overall attitude.

5 Conclusion

Public engagement is key in the development of innovative and technological advances. Especially in the medical field communication between professionals and the general public is essential in order to implement new techniques and treatment that would provide better patient care. This study presented a survey, in order to measure the European attitude towards public engagement in the field of personalized medicine.

The results show that the young highly educated European audience shows strong public support for the field of personalized medicine. Compared to an earlier American study, they feel likewise excited but unfamiliar with the subject. Investing in public engagement could help answer their questions and understand their concerns. As the communication between public and researcher in this field is particularly vital it could aid both the implementation and the development of new techniques.

Future research should be done to identify the attitude of the whole European public and the public engagement strategies that could be applied

References

1. RRI tools:Home Page - RRI Tools. https://www.rri-tools.eu/es .
2. Y. Bombard, J. Abelson, D. Simeonov, and F. P. Gauvin:Citizens' perspectives on personalized medicine: A qualitative public deliberation study, Eur. J. Hum. Genet., vol. 21, no. 11, pp. 1197–1201, (Nov. 2013).
3. V. J. Dzau and G. S. Ginsburg:Realizing the full potential of precision medicine inhealth and health care, JAMA - Journal of the American Medical Association, vol. 316, no. 16. American Medical Association, pp. 1659–1660, (2016).
4. L. Towart:Personalized medicine: Top Four Misconceptions - BioTechniques, (2018). https://www.biotechniques.com/precision-medicine/the-top-four-misconceptions-of-precision-medicine/ .
5. KRC Research:Public Perspectives on Personalized Medicine A Survey of U.S. Public Opinion, no. May, p. 16, (2018).
6. Formularios de Google. https://docs.google.com/forms/u/0/..
7. Microsoft Excel. https://support.office.com/ .
8. R. M. Groves:Survey methodology. J. Wiley, 2004.
9. P. Lavrakas:Encyclopedia of Survey Research Methods. Sage Publications, Inc., 2012.
10. Validity and reliability of questionnaires. https://es.slideshare.net/Venkitachalam/validity-and-reliability-of-questionnaires.

Exploring the Application of Research on Responsible Artificial Intelligence Over Time

[a]Dougal, Shakespeare,
dougalian.shakespeare01@estudiant.upf.edu
[b]Alexander Keijser, UPF,
alexandernicolai.keijser01@estudiant.upf.edu
[c]Francesca Ronchini, UPF,
francesca.ronchini01@estudiant.upf.edu [d]Pavlo
Apisov, UPF, pavlo.apisov01@estudiant.upf.edu

Abstract: The field of Artificial intelligence (AI) has recently seen a paradigm shift from the traditional symbolic approach towards system performance enhancements resulting in new algorithms / technologies for which the liability of a software's deviation is ambiguous. Researchers have acted to tackle technical and moral AI issues that may arise through publications highlighting the importance of responsible research in AI. Our research assesses the extent to which such research is being practi- cally considered in state of the art AI research. We quantify usage as a citation to the publication under consideration. We have considered four of the top institutions publications on AI safety and responsible research from the period of 2009 - 2019 and analysed correlations strengths in the time domain. Our results do not find any significant correlations which we theorise is due to weaknesses in data selection strategies. For future developments, we suggest increasing the number of papers analysed and considering only technical papers that cite AI safety research.

Keyword: Artificial Intelligence, AI Ethics, Responsible Research, Public Engagement

1. Introduction

The behavior of the future machine has become less predictable in recent years since the development of autonomous genetic algorithms, learning machines and agent ar- chitectures has become more prevalent. Modern machines are now capable of au- tonomous decision making, and in some cases, can act without human involvement (adaptive capabilities) posing issues for both researchers, policy makers and consumers alike.

In research, a paradigm shift in the field has occurred. The traditional symbolic approach of Artificial Intelligence (AI) has, although still an active research area, seen a conversion in focus towards system performance enhancements. This has lead to the dispersing of program flow control resulting in data becoming less easily and directly interpreted. Moreover, this has resulted in an increasing number of machine actions for which no party is able to assume responsibility, from both an ethical and technological stand point [8]. With AI researchers in the field having now proclaiming there is a chance of AI outperforming humans in all tasks within this century [5], public concern for both social welfare and technological issues is becoming a more

prominent topic of public discussion and must subsequently be addressed by the academic community.

Researchers have responded to these concerns through a multidisciplinary approach in order to solve the problem of aligning AI systems with human values [6]. Specif- ically, the involvement of social scientists has been deployed to aid the understand- ing of human cognition, behavior, and ethics with journals such as Minds and Ma- chines[9] publishing a consistent publication of papers focusing on the ethical impli- cations of AI research. Technical considerations have also been researched thoroughly through the likes of DeepMind[3], OpenAI[10] and the Center for Human-Compatible AI (CHAI)[2] with increasing research being published in domain specific research groups focusing on AI safety. AI safety is a research field defined to be concerned with *'mitigating accident risk [...] in terms of classic methods in machine learning, such as supervised classification and reinforcement learning'*[1]. Typically, this type of research takes a more technological approach to resolving AI accident risks.

Our research will aim to determine the impact of such research on responsible prac- tices in AI within the field of academic research (mainly focused on technological applications but we will discuss other research fields). We will conduct meta research to explore the extent to which research on responsible AI usage and development is being applied in the field. We justify the importance of this research as AI ethics re- search has seen an increasing focus on practical application in recent years and thus, it is important to assess if these advancements are also being incorporated into modern state of the art systems and assess if there has been any deviations or developments over time. We can justify our researches originality as it is not only assessing historic literature on the field of research on responsible AI usage, but also state of the art research for which correlations have not yet previously been identified or explored.

2. Research Methodology

For the purpose of our research we define citation count as a metric to measure the application of responsible research focusing on the field of AI. We use this metric to subsequently assess the extent to which research of such nature is being both considered and applied in (mainly technical) academic literature.

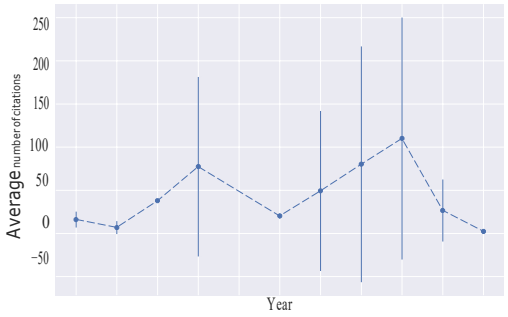
Our research focuses on exploring the application of publications from domain specific influential journals and organizations that specialize in responsible AI research. In particular, the institutions considered are DeepMind, OpenAI, the Center for Human- Compatible Artificial Intelligence (CHAI), and the Machine Intelligence Research In- stitute (MIRI)[7]. We chose these organisations for their high publication count and associated researchers high credibility in the field, hence making them strong candi- dates to assesses the extend to which their research is being applied by researchers in practice.

The next stage of our research was to select from our chosen institutes / organisations, papers falling under the classification of AI safety. From these results, we compute for each year in the period 2009 - 2019: the total number of publications per organi- sation, average number of citations for all organisations, total number of citations per organisation and average number of citations per organisation. These findings can be found in the Results section 3 of this document.

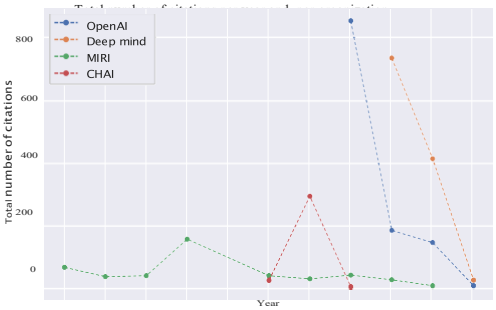
3. Results

In order to get insight into the development in time of the impact of AI safety lit- erature, we visualise several statistics about publications on AI safety by *OpenAI*, *DeepMind*, *MIRI* and *CHAI*.

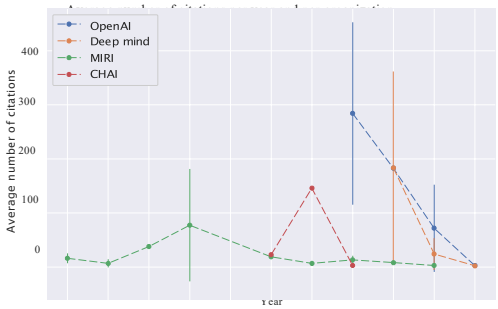
(a) Plot showing the number of publications per year for each of the organizations.



(b) Plot showing the average number of citations per publication/paper per year for all organiza- tions combined. Error bars indicate a single stan- dard deviation interval.



(c) Plot showing the total number of citations per year for each of the organizations.



(d) Plot showing the average number of citations per publication/paper per year for each of the or- ganizations. Error bars indicate a single standard deviation interval.

Figure 1: Plots showing statistics of the AI safety papers published between 2009-2019 by the organizations *OpenAI*, *DeepMind*, *MIRI* and *CHAI*

Figure 1a shows the number of publications per year for each of the organizations. We can see that the number of publications by *MIRI* is relatively steady, the other organizations started later, with *DeepMind* publishing significantly more papers in 2018 than in 2017 and 2019. It must be noted that a lower number of publications from 2019 is to be expected, as the data used for this study was obtained before the end of 2019.

Figure 1b shows the average number of citations per publication for all organizations combined. This should give an indication of the development in time of the impact of responsible research in AI. The average seems to follow a kind of trend: it increases between 2010 and 2012, decreases between 2012 and 2014, increases again between 2014 and 2017 and decreases from 2017 onward. However, as the standard deviations in the number of citations per year are very large, there is too much uncertainty to draw any conclusions.

Figures 1c and 1d show information of the numbers of citations per organization. We include both the total and average per year to give the most complete picture. Figure 1c shows the total total number of citations per year for each of the organizations.

This gives some information about the impact of the different organization but might not be very informative in isolation, as this number is directly related to the number of publications. The importance of including multiple different statistics can be illus- trated with an example: we can see that *DeepMind*'s publications from 2017 have the most citations, although significantly more papers were published in 2018 (see figure 1a). *MIRI* is the most stable in the number of citations, apart from a slight increase in 2012, also in terms of the yearly publications (figure 1a). Our data only contains relevant publications by *CHAI* from 2014, 2015 and 2016, with a peak in the number of citations occurring in 2015. Both *OpenAI* and *DeepMind* show a decrease in the number of citations after their first year.
Figure 1d shows the yearly average number of citations per publication for each of the organizations. This provides more detailed information than the overall numbers in figure 1b. For instance, we can see that usually when a peak occurs, the standard deviation becomes larger as well. This means that the average was raised by a small subset of all publications of that year that were highly cited. The increase might even have been caused by a single very highly cited paper. The average numbers in 1d show more or less the same trends as the totals in 1c.

4. Discussion

Overall, the results do not show us any interesting highlight regarding the data which would allow us to get into concrete conclusions. We believe that using the citation year instead of publication year would lead to more significant conclusions, especially regarding correlations in the time domain. To give a more concrete example - our data have the dates of when AI safety papers were published and we correlate their citations with these years, whereas in reality most of the citing papers were published in the future and in different years. Thus, for instance, a paper from 2016 could affect 2019 more than previous years. Another important issue is using the number of all citing papers instead of only technical that cite AI safety research works. We discuss our decision for quantifying the number of all citing papers in the limitations section (4.2).

4.1. Data

We also understand that, taking into account the size of our of data (79 papers), it is not possible to generalize or to find a well defined visible trends which would permit us to understand and interpret better the results. Finding the right size of dataset and the data itself in this case is not an easy task. The methodology of scraping academic search engines such as Google Scholar and Scopus to gather more papers relating to responsible research in AI was considered in early research stages, however. From preliminary experiments it was deemed that this methodology offered little guarantee of only obtaining AI safety-relevant papers and thus was not a reliable technique to be implemented.

4.2. Limitations

With the proposed above data acquisition methodology we also foresaw other two problems which could arise in the data collection stages of research: (i) selection bias and (ii) treatment differ. We address the problem of selection bias by considering only papers which have been explicitly tagged by by the organizations as falling under the category of AI safety. Thereby literature selection becomes an automated process exempt from human bias which could otherwise be imposed.

In the case of treatment differ, our proposed methodology suffers from the issue of categorising citing papers as technical or non-technical by our own subjective judg- ment. To make categorisation less biased and further increase objectivity it would be necessary to seek further guidance from those with additional technical expertise in the AI field, a development which could be implemented in future work. It is worth noting that for many papers, categories are clearly defined through the nature of re- search (for example, [11] is clearly defined to be theoretical and [4] to explore more technological aspects) and hence it may only be necessary to deploy expert domain specific knowledge in cases which lie on the periphery of technical and non technical.

We resolve this issue in our research by only quantifying the number of citing papers from different organizations without any reference to the technical or non-technical nature of the papers. This way we can assess the influence of the selected research groups on developing responsible AI systems from a wider perspective. We believe this makes our research more reliable but unfortunately less insightful when applied to more specific technical or non-technical research domains.

5. Conclusions

The study had the aim to evaluate to what extent responsible research have been con- sidered and applied on technical academic literature on the field of AI. We considered the number of citation to the papers under analysis as our metric to measure practical application of responsible research, analyzing correlation between the time domain. We restricted our research to four institutions focusing on AI safety and responsible research (DeepMind, OpenAI, CHAI, and MIRI) between the years 2009 and 2019.

The results convey that the number of publications per year per organization vary according to the organization. They show a stable trend for MIRI, CHAI and OpenAI, meanwhile the number of publications from DeepMind steeply increases from 2017 to 2018. The average number of citations per publication for all organizations combined seem to follow a trend increasing and decreasing every two years but it is not possible to get certain conclusion due to the high standard deviations. We also considered the total and the average of number of citation per year of each organization. It is interesting to notice that the most cited publications from DeepMind are in the 2017, although the same organization significantly published more papers in 2018. OpenAI and DeepMind show a dramatic decrease in the number of citations after their first year of publications. From the data, it is notable a correlation between the peak in average number of citation and the standard deviation. This lead to conclude that the average was raised only by a small subset of all publications of that year which were highly cited.

Finally, it was not possible to draw any concrete conclusion for several reasons: (i) the methodology suffers from the issue of categorising citing papers as technical or non- technical, (ii) small amount of data, (iii) using citation year instead of publication year would probably lead to more meaningful results. Anyway, we think that the results and conclusions coming from this research could be considered as preliminary for future analysis. Future works should consider all the previous mentioned limitations which should bring to more reliable and meaningful results.

References

[1] Dario Amodei et al. “Concrete Problems in AI Safety”. In: (2016), pp. 1–29. arXiv: 1606.06565. URL: http://arxiv.org/abs/1606.06565.

[2] *Center for Human-Compatible Artificial Intelligence*. Accessed: 2019-11-10. URL: https://humancompatible.ai/.

[3] *DeepMind*. Accessed: 2019-11-10. URL: https://deepmind.com/.

[4] Krishnamurthy Dvijotham et al. "A Dual Approach to Scalable Verification of Deep Networks". In: *UAI*. 2018.

[5] Katja Grace et al. "Viewpoint: When will ai exceed human performance? Evidence from ai experts". In: *Journal of Artificial Intelligence Research* 62 (2018), pp. 729–754. ISSN: 10769757. DOI: 10.1613/jair.1.11222. arXiv: 1705.08807.

[6] Geoffrey Irving and Amanda Askell. "AI Safety Needs Social Scientists". In: *Distill* 4.2 (2019). ISSN: 2476-0757. DOI: 10.23915/distill.00014. URL: https://distill.pub/2019/safety-needs-social-scientists.

[7] *Machine Intelligence Research Institute*. Accessed: 2019-11-10. URL: https://intelligence.org/.

[8] Andreas Matthias. *The responsibility gap: Ascribing responsibility for the actions of learning automata*. Tech. rep.

[9] *Minds and Machines*. Accessed: 2019-11-10. URL: https://www.springer.com/journal/11023.

[10] *OpenAI*. Accessed: 2019-11-10. URL: https://openai.com/.

[11] Carl Shulman et al. "Machine Ethics and Superintelligence". In: Machine Intelligence Research Institute. 2009.

The rigor-relevance debate in the context of dissertation topic selection

Darius Petermann, Ignasi Nou, Christian Steinmetz, and Myrsini Ioannou

Master in Sound and Music Computing
Universitat Pompeu Fabra, Barcelona,
Spain
{dariusarthur.petermann, ignasi.nou, christianjames.steinmetz, myrsini.ioannou}01@estudiant.upf.edu

Abstract. The dissertation topic selection process represents a crucial step in the journey of an aspiring doctoral student, and consequently many factors are to be taken into consideration. One of these considerations requires the examination of the dichotomy between rigor and relevance in the context of academic research. In this paper, we first ex- plore this problem from a broader perspective by presenting the various challenges that developing a dissertation topic entails. We will then fo- cus more specifically on the issue of rigor and relevance in the context of doctoral studies. Through out examination we hope to bring to light the value of balancing these two aspects and in the process we propose different methodologies to address this imbalance.

1 Introduction

1.1 Rigor-relevance debate

Throughout nearly all fields of academic research there exists a common criticism of the apparent irrelevance of research output with regards to its application in industry and society more broadly [15]. This issue is often framed within the context of what is known as the rigor-relevance debate, wherein research output is evaluated by its academic rigor and relevance to the greater field (e.g. industry, society, etc.) [5]. Different models exist for these characteristics, either where the two are directly opposing each other and another where they are in fact orthogonal. In most cases, where a research work falls in this spectrum defines whether it will be published in academic journals, and an example of this for the orthogonal model is demonstrated in Figure 1. This is also closely related to the so called theory–practice divide, or scholar–practitioner divide, all of which are focused on examining the challenges involved in the application or transfer of knowledge from academic research to industry and society as a whole [9].

Traditionally, academia has produced work more rigorous in nature, with a focus on the formulation and assessment of theoretical foundations, often without direct regard for its application. The motivation to undertake rigorous research is closely tied to the process of establishing one's academic prowess, in addition to attaining publications in peer-reviewed journals, academic outcomes that both

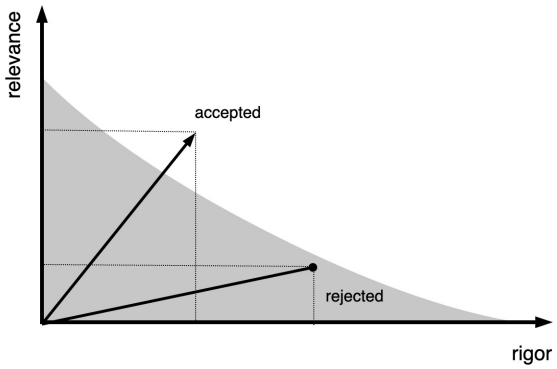


Fig. 1. Depiction of rigor-relevance within the context of academic publishing

serve as primary metrics with which academics are judged [31]. These realities only further serve to incentivize academics to pursue such work with little regard for industrial relevance.

In contrast, more recently, there has been a growing interest in increasing the relevance of academic research output with regards to its application within industry. This is largely influenced by the ever increasing role of industrial fund- ing in academic research. In 2007, the majority of all academic funding within the United States was supplied by private sources [11]. This reality, coupled with the dwindling number of tenure track faculty and researcher positions at universities, often pressures researchers to undertake research in a manner that emphasizes its value within an industrial context, and potentially at the cost of academic rigor from the perspective of some scholarly critics [8].

1.2 Challenges in topic selection

The selection of a dissertaion topic is one of the most difficult and significant de- cisions a new graduate student can make. Although this is a personal decision, the selection process must be based thoughtfully on solid foundations, deter- mined not only by personal preferences, but also by objective data that reflect academic reality.

Research has shown that often topic selection is not an independent choice of the student, but it is assigned by an advisor [21]. Depending on the discipline, the independence varies. The physical sciences and engineering have the lowest amount of independence, in contrast to social sciences and humanities that had the highest. In addition, studies showed that there are several factors that affect the topic selection. Based on student's criteria, the factors with the highest pri- ority are the familiarity of the subject, the current trends in the field and their own life experiences. Also, other significant factors are adviser's preferences, like- lihood of publication, and job prospects. [12]. Even though the factors mentioned are valid, there might be several challenges that may make it difficult for the

student to complete the dissertation or create obstacles in conducting unbiased research. For example, choosing a dissertation topic based on personal interests or "passion" might have implications regarding emotional relevance that leads to confirmation bias [26]. An emotionally involved researcher might have a bias in data selection and interpretation or might search for evidence in convenient places. Furthermore, students must pay attention to the influence of current trends on their choice of dissertation topic. Despite the fact that following a trend might provide sufficient quantity of resources or a feeling of competence for being part of something bigger, it's possibly more challenging to be seen and also there is no guarantee that the momentum of the topic will be the same after the completion of the dissertation [26].

Besides the challenges resulting from the students' criteria, there are also other crucial factors that are often neglected and must influence a student's choice of the dissertation. For instance, feasibility and the availability and qual- ity of data is extremely important when conducting research. These may affect the general credibility and reproducibility of the research [26]. Moreover, one of the most essential factors that sometimes is ignored and is the core purpose of research is the contribution to scientific knowledge [30]. Additionally, the dis- sertation topic must be chosen according to the future ambitions of the student since it will undoubtedly be important in the search for employment. These are just some of the challenges the doctoral students face more broadly in the disser- tation topic selection process, and in the next section we will examine existing strategies in the literature for addressing some of these challenges.

1.3 Existing strategies for topic selection

As outlined in the previous section, the process of choosing an ideal dissertation topic is complex, involving many different factors, that both students and their advisors, must consider [17]. Moreover, we have to take into account that this process can be stressful and time-consuming [25]. Several works in the literature have already presented and examined different strategies that can simplify this process. In particular, based on these papers [30] [12], we provide a brief de- scription of existing recommended strategies in the dissertation topic selection process.

Use of advisors, professors, and scholars Established members of academia often have much insight to lend to the doctoral student as they are searching for a topic. In this case a student can follow one of two paths. The most common is for students to provide a general idea of their area of interest and ask potential advisors if they have any proposals prepared that follow this area. In addition, the student may first assemble a collection of topics themselves and then consult with possible advisors to receive feedback. In many academic settings it is the advisors who propose part of their research as a topic of choice [24].

Study of relevant literature The literature that the student shows interest in can also be a good method of identifying a relevant topic. Students should notice

what topics are commonly addressed and potential areas for continuation of these works. They should also note the structure of published theses and dissertations, as well as the names of the researchers serving on the advisory committee. They should also pay attention to topics that interest them, including the reference and literature review sections. These sections may allow students to develop their own unique research ideas and designs, that are of interest to existing researchers in the field.

Curriculum and conferences The graduate school curriculum can be also an effective tool when searching for ideas for possible topics [17]. On the other hand: faculty and student factors, nature of a topic, trend, duration of study, research funding, eventual audience citation, are some examples of factors that can influence when choosing the topic that we previously found [24]. In fact, strategies and factors during the selection or search for a topic plays an important role in the balance of rigor and relevance.

1.4 Goals of doctoral students

As has been mentioned previously, a crucial factor to take into consideration while investigating the dissertation topic selection process are the post-doctoral goals of the aspiring student. Depending on where the students picture them- selves in their post-doctoral journey - that is, either working in academia or in the industry sector - the desired balance of rigor and relevance involved in approaching a topic will vary drastically.

In 2017, Nature led a comprehensive survey [34] involving more than 5,700 PhD candidates, from all around the world. In the survey, the respondents were asked various questions regarding their experience as doctoral students, including the sector they were seeking to pursue a career in. The results showed that more than 50% of the subjects were hoping to secure an academic job, while 22% of them wished to pursue an industrial career path. Another survey led by Nature a few years early [33], which investigated the problem of academic expectations among graduate students, indicated that more than 50% of the respondents did not foresee a future in academia, as the field was too competitive. It is thus clear that not all PhD candidates are seeking an academic career. Moreover, according to a recent publication exploring the lack of job availability in academia [16], only 12% of doctoral graduates attain academic positions in the USA. In reality, doctoral graduates tend then to turn down their initial aspiration in academia to seek industry opportunities instead.

Regardless of the student's desired career path, an academic institution should always be able to provide a proper supervision adapted to the student's needs, which is often not the case. The Nature survey [34] revealed that an im- portant portion of the respondents felt that their supervisors were not open to the idea of them pursuing a career path outside academia, nor were they pro- viding useful career-related advice. These results depict an overall substantial reluctance from faculty in preparing their students for an industrial career. In

her article depicting the biased training that PhD candidates receive through their academic journey [2], Sarah Anderson stresses the fact that doctoral pro- grams are designed to accommodate for students seeking to remain in academia. This introduces a bias in the selection of dissertation topics towards topics that display a greater level of rigor, and less relevance, independent of the aspirations and goals of the student.

2 Considerations of rigor and relevance for thesis topics

2.1 Case for rigor in dissertation topics

We must focus on the aspects and characteristics of the role of rigor and rel- evance, and in this section we will begin by addressing rigor. But, why is it important to be rigorous both during the elaboration of a thesis and what role does the dissertation topic have? The main purpose of rigor in research consists in being able to generate solid, replicable, and stable claims and results. The rigor of these results will establish confidence among the scientific community, allowing progress in the field of research.

Concerning the research process itself, this must be strict in applying the scientific method to ensure a robust and unbiased study. It is also necessary to take into account the different parts of the work, and to ensure robustness and impartiality in the methodology, analysis, interpretation, and elaboration of the results. Work following this criteria will contribute to the rigor of the research, as well as to the transparency of the experiment and will facilitate its reproduction and extension [10].

Those types of research that are apparently rigorously justified, i.e., ques- tioning the deliberate falsification, fabrication, proposal plagiarism, execution or processing of results; will be considered rigorous of a misleading or insidious type. At this end we could consider this lack of rigor as a lack of "Ethical Rigor", since it is unlikely that such lack of rigor will be caused involuntarily. Ethics must be a very important factor to take into account during the development of our research [19]. Therefore, a minimum of rigor must be also considered.

Research in which the scope of low rigor is no longer voluntary and leads to false or erroneous conclusions will be categorized as "creative rigor". This type of lack of rigor will be more notorious in situations where, for example, data is selectively chosen or no significant results are shown to support the hypothesis. It will be important to achieve a level of creative rigor if our objective is the independent reproducibility of our research. It will also allow us to be consistent with our original hypothesis and obtain honest results.

A next level of rigor is defined as the one in which the researcher applies rigor where "it is easy to apply", both by knowledge and by urgency. It is probably the most common behavior, since modern research currently involves a considerable level of urgency in timelines; rigor will be applied to those points that are easier for us to justify. Reaching this level "careless level" of rigor, will reinforce the objectivity of our research and therefore will keep us in the right direction of both the defined research method and the hypothesis.

A next step of rigor enforcements takes into account the researcher's criteria. The researcher will decide on which aspects or sections has to apply rigor. At this point, the rigor level will be significant, reaching this level will imply justifying all those aspects that are more or less difficult and we will read them as outstanding or relevant.

Related to the previous point, we observe that the lack of rigorousness, both in uncomplicated parts as well as in the parts that will not be taken into account (based on our point of view), will be important during our doctoral thesis. In the end, this type of scientific rigor practice allows the researcher to select where rigor might be inappropriate, leading to ambiguous results and a reduced probability of reproducibility.

However, rigor is now required, and journals, publishers and editors have established a set of standards for authors, ranging from basic guidelines to ro- bust ones. While such guidelines aim at transparency or scientific rigor, it is ultimately to focus on the researcher profile, the one responsible for conducting the experiments in an ethical and transparent direction. Such methods are es- pecially necessary to understand the importance of scientific rigor and integrity when a hypothesis is refuted by rigorous science. Consequently, addressing these topics (ethical and transparent direction) will not only improve the probability of reproducing the results of experiments, but will also increase the probability of independent reproducibility between investigations, leading to long-term rigor and improving scientific integrity.

Finally, highly rigorous research will play an important role in tackling the term ”reproducibility crisis”. These are the researches that will be subjected to reproducibility studies. While the rigor reaches a very high level, there are unknown variables and conditions that will question the accuracy of repeated results. However, it should always be noted that, without a high rigor baseline, the replication of experiments will be much more complicated, if not impossible. This high rigor type of research (and the other ones detailed above) should be taken into consideration during the research (and selection) of our doctoral study.

During our topic selection we should take into account that certain level of rigor, the ones explained above, must be considered both if we want to reinforce the theoretical part of our practical application and if we want to be considered in those academic publication resources in which rigor is now required.

Now rigor is an important issue to tackle if our research is related to prac- tical applications. Moreover, regarding the topic selection, we should be able to demonstrate (to a certain level) the rigorousness of our research of the topic selected. But we will observe that rigor is not always mandatory and expected in relevance topics (e.g. practical application), when, as explained above, should be.

2.2 Case for relevance in thesis topics

As addressed in Section 1.4, it is crucial for students to take into consideration their post-doctoral aspirations while in the midst of selecting their dissertation

topic. For students seeking to pursue a corporate or industrial career after their doctoral journey, it may be in their best interest to connect their research work to practical applications. As described in [7], a doctoral degree has the potential to be used as a valuable asset in any discipline outside of academia. Beside being a tangible proof that the student mastered a certain topic to its highest academic degree, having the dissertation topic closely related to an industrial application can demonstrate the doctoral student's ability to connect their academic research to concrete and practical applications of it. Notably, research involving a rigorous approach, while demonstrating little to no practical applications, may be found as a less powerful asset in securing an industrial position.

According to a 2010 Royal Society report [27], the vast majority of people undertaking a doctorate will ultimately end up working outside of academia. The journey of a doctoral student, from early research career to full professor, is accentuated by key transition points at which the percentages of academi- cians will drop substantially. Of 200 doctoral students, only 7 will be offered a permanent academic position [32].

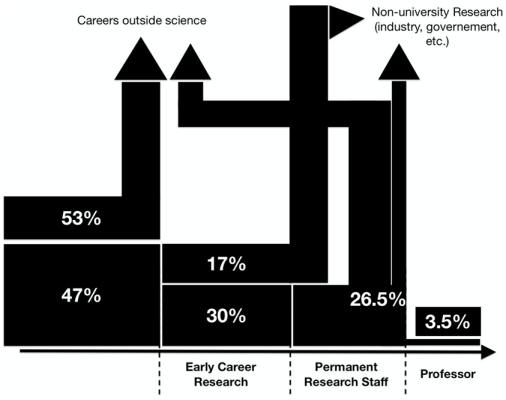


Fig. 2. Flow of scientifically-trained PhD candidates into other sectors [27]

Focusing a research work solely on rigor while ignoring relevance is thus often not in the best interest of most doctoral students. From a purely professional perspective, adding a certain degree of practicality to a research work can only be beneficial, especially if the student is planning on pursuing an industrial career path after their studies.

In their article on the impact of research outside of the academic sphere, Biswas et al. stress the fact that 82% of articles published in scientific journals are not even cited once while an even larger portion never reach the popular media [3]. This visibility issue means that research works with great practical potential would probably remain within academia and never reach practition-

ers. Stepping beyond academia by increasing relevance in a dissertation can thus carry its impact potential outside of the academic scope. By making doctoral research more accessible to practitioners, students could substantially improve their thesis' impact towards practical applications. Such an approach could ulti- mately lead to a better exposure of the student's research beyond academia and, ultimately, lead to an increase in industrial innovation and economic growth [13]. While many aspiring doctoral students are aiming at a long-term academic career, this study [27] has shown that most will end up pursuing a career in the industrial field, either by choice or due to lack of opening in academia. From this perspective, maintaining a certain level of relevance in a thesis could tremendously help these students to build their future career beyond academia. Statistics aside, carrying a fair balance between rigor and relevance will also allow the research work for more visibility outside the academic sphere, which could result in expending the student's network beyond academia. Nevertheless, balancing rigor and relevance in the context of academic research is still a widely debated subject and increasing relevance is not always an easily achieved task.

3 Reconciling the rigor-relevance debate in doctoral studies

3.1 Addressing relevance in thesis topic selection

Some scholars argue that science should be in distance from practice in order to maintain its integrity. Others indicate the impossibility of consolidation between rigor and relevance [23]. However, scholars who support the requirement of aca- demic relevance in doctoral studies recommend the involvement of practitioners [4]. Researchers and practitioners tend to work in isolated groups with rare co- operation but undoubtedly the contrary leads to the advancement of knowledge for both cases [23]. A doctoral student must take into account the advantage of efficiency and broader understanding that relevance provides, especially when deciding the topic of his or her research.

Without a doubt, relevance is desired in research but measuring it is challeng- ing. On the other hand, rigor can be assessed more easily [29]. This is the reason that the academic community often shows a preference for rigor. Students must take into consideration the following issues in order to select a relevant topic.

A doctoral student must identify "real world problems" in order to assure relevance. By this identification the likelihood of practical applications is higher as well as the provision of new useful insights. The choice of the topic can be specified by answering the following questions:

1. Are the research questions relevant to practitioners? [23] [29]
2. Can the research idea be generalized to other settings? [23]
3. What are the possible audiences that would listen to this research? [29]
4. Can the research question be answered rigorously? [29]

Develop networks Students can develop networks in order to integrate their effectiveness, especially at the early stages of a student's research where the topic may not have been particularly developed or identified. By maintaining rigor and simultaneously collaborate with fellow students with corporate or practical experience a doctoral student can increase relevance [22] and at the same time efficiency. The insights that experienced students can convey can contribute significantly and lead to a better identification of the topic in terms of relevance.

Invite practitioners on-campus Inviting practitioners to speak to student groups and conferences on-campus, or attending practitioners' conferences [28], is an opportunity that promotes the interaction between scholars and practitioners. This leads to a better understanding and development of a practical perspective. Academics rarely attend practitioner conferences that often provide excellent insights about the best practices, challenges, and solutions that are currently relevant [28]. Moreover, organizing crossover workshops is a dynamic way for scholars to present their progress to professionals interested in academic research as well as to network and find out the topics that resonate more with practitioners [28]. This interchange has great significance especially at the initial point of a student's research since topic selection or specification can be influenced by practitioners' insights.

Field visits and working as a practitioner Unquestionably, conducting field visits can help doctoral students acquire practical knowledge by observ- ing professionals in practice. This involves analyzing what challenges they en- counter and which solutions they implement. Practitioners must be approached as beneficiaries and prospective recipients of research. They must be convinced of its potential. Through this effort, a communication is established between re- searchers and practitioners, and relevant research problems are identified more effortlessly. These problems can form the topic of a doctoral student's research. An even more effective method is for scholars to work as practitioners in order to perceive the tangible challenges of practice [28]. In addition, industry con- sulting or attending board appointments and executive education is regularly a sufficient way to acquire practical knowledge that ensures relevance and will lead to a more relevant topic selection. [4]. Generally, practical experience is very likely to motivate doctoral students to choose a relevant topic that could have practical applications and practitioners could also take advantage of.

Convey relevant insights and collaborate with practitioners According to some authors, doctoral students can benefit by conveying relevant results to practitioners. Getting feedback from questions and comments can serve as a tool to specify more precisely the research topic and lead to a relevant orientation [28]. This can be achieved in multiple ways. First, by presenting at practitioner conferences [28]. This direct method provides an opportunity to bridge the di- vide between academics and practitioners, and usually is neglected by the former.

Secondly, another way of reaching practitioners is to write for crossover journals, trade press, or newspaper op-eds [28]. The goal of this is to communicate in an interesting and engaging way the current discoveries in the field of research that the scholar is involved. Even though the process may demand serious effort and might be time consuming, conveying the findings to practitioners is extremely meaningful in terms of receiving response. Finally, collaborating with practi- tioners by coauthoring can supply researchers with information that might be inaccessible in any other case [28]. This kind of partnerships can portray subtle interpretations that can be helpful to doctoral students in order to select a more relevant topic or shift it to a more relevant direction.

3.2 Addressing rigor in thesis topic selection

To begin to address rigor within the context of doctor studies, we must first examine the purpose of the doctoral thesis, and the role rigor plays. At its core, the purpose of a doctoral degree is to demonstrate the student's ability to carry out novel research independently, and present these results formally to other researchers in their field [6]. Clearly a core principle of a successfully defended thesis is the application of sufficient rigor, but what constitutes sufficient rigor is in itself a subject of contention among institutions, which have varying standards and procedures [1]. While it is certain that a successful doctoral thesis will be rigorous, there is room for the doctoral student to decide how to balance rigor and relevance, as has been introduced in previous sections. In this section however, we aim to provide insight into methodologies for achieving greater rigour within the thesis topic selection process.

While a thesis topic in of itself does not generally directly indicate the rigor of a research work, the overall framing of the topic can lend itself to a path of greater rigor. For our discussion here, we will consider what we have named the level of "achievable rigor", or in other words, the feasibility of achieving a level of rigor, given the resources available during the completion of a doctoral thesis, namely the time constraint [18].

Determining research topic scope The first factor to address is the scope of the proposed topic. In this case, scope refers to the complexity of the research questions that will be investigated in the context of the thesis. Most doctoral thesis candidates spend somewhere between three to seven years pursuing re- search in preparation for their defense. This means that the scope of the thesis topic will in some way impact the level of achievable rigor. By reducing the scope of the proposal, this provides the opportunity for deeper investigations and po- tentially the ability to achieve greater rigor. Carefully setting the scope for the thesis topic is one of the most powerful ways to address the level of rigor in a doctoral thesis, as it often does not directly impact the relevance of the research, and in some cases, it may even increase the relevance.

Incremental vs. innovative research Next, we must also consider how the proposed topic fits within the body of existing research. While potentially not

apparent, the extent to which a proposed topic extends or deviates from the body of previous work can have a significant impact on the level of achievable rigor within the doctoral thesis. In general, as research in a field progresses, the body of knowledge in this area, in most cases, is broadened and further corroborated. This leads to a compounding of knowledge, wherein incremental advancements are achieved that ultimately result in significant evolution within the field over time. This is a manifestation of the famous metaphor of present day scientists "standing on the shoulders of giants", an often used phrase to express the reality that our progress and advancement is directly built upon the work of those who came before [14].

To that effect, during the topic selection process it is important to consider in what way the proposed topic extends existing research in the field. Broadly, research has been examined within a framework of incremental and innovative contributions [20]. The degree to which this proposal is a clear extension or incre- mental in nature, brings about a proportional increase in the level of achievable rigor, as the existing work provides a convenient framework, which inherently provides rigor to the work. This more greatly aids in achieving rigor since the researcher need not independently develop and defend each component of their work. For this very reason, most doctoral research topics follow this path of mak- ing focused contributions to well established fields, and achieve sufficient rigor doing so. The contrasting situation involves topics that aim to investigate new or undeveloped areas within a field, potentially bringing about significant inno- vation. While often promising, these topics pose a greater challenge in achieving the same level of relative rigor, since much more extensive work must be carried out to validate and defend the claims and findings, in comparison. In such a case, properly defining a sufficiently narrow scope of the research questions addressed is critical to a successful research result within the time constraint of a doctoral program.

Ultimately, there is only so much that can be done during the phase of topic selection to directly increase the rigor of doctoral research. Nevertheless, from our examination in this section, it is clear that some considerations made early on in the process of topic selection have an impact on the level of achievable rigor. Addressing and achieving an appropriate balance between both the scope and apparent incongruity of the proposed topic can provide a clear path for the doctoral student in their journey of executing rigorous research.

4 Conclusion

The debate between the role of rigor and relevance will continue to persist in academia, and as along as it does, this factor remains an important consider- ation for researchers. While doctoral students are often inundated with many different external factors during the determination of their thesis topic, the bal- ance between rigor and relevance is often less examined. In this work we outlined the factors at play in this debate within academia in general, and then made a case for both in terms of academic research. Finally we presented potential

methodologies for doctoral students to adopt in order to better address this de- bate within the process of selecting their dissertation such as the involvement of practitioners when addressing relevance or the considerations on topic selection regarding the level of achievable rigor when addressing rigor. We hope that this work elucidates the benefits of achieving balance between rigor and relevance and may provide an additional perspective for doctoral students as they are selecting their dissertation topic.

References

1. The past, present and future of the phd thesis. Nature **535**(7610), 7 (July 2016). https://doi.org/10.1038/535007a, https://doi.org/10.1038/535007a
2. Anderson, S.: Make science PhDs more than just a training path for academia. Nature **573** (2019). https://doi.org/10.1038/d41586-019-02586-5
3. Biswas, A.K., Julian, K.: Prof, no one is reading you. Straits Times (2015)
4. Finch, D., Falkenberg, L., McLaren, P., Rondeau, K., O'Reilly, N.: The rigour–relevance gap in professional programmes: Bridging the 'unbridgeable' be- tween higher education and practice. Industry and Higher Education pp. 152–168 (04 2018). https://doi.org/10.1177/0950422218768205
5. Finch, D., Falkenberg, L., McLaren, P.G., Rondeau, K.V., O'Reilly, N.: The rigour– relevance gap in professional programmes: bridging the 'unbridgeable'between higher education and practice. Industry and Higher Education **32**(3), 152–168 (2018)
6. Gould, J.: What's the point of the phd thesis? Nature News **535**(7610), 26 (2016)
7. Hankel, I.: Why earning a phd is an advantage in today's industry job market. Nature (2019). https://doi.org/10.1038/d41586-019-00097-x
8. Hart, K.: Is academic freedom bad for business? Bulletin of the Atomic Scientists **45**(3), 28–34 (1989). https://doi.org/10.1080/00963402.1989.11459664, https://doi.org/10.1080/00963402.1989.11459664
9. Heracleous, L.: Introduction to the special issue on bridging the scholar–practitioner divide (2011)
10. Hofseth, L.J.: Getting rigorous with scientific rigor. Carcinogenesis **39**(1), 21–25 (jan 2018). https://doi.org/10.1093/carcin/bgx085
11. Hottenrott, H., Thorwarth, S.: Industry funding of university research and scientific productivity. Kyklos **64**(4), 534–555 (2011)
12. Isaac, P.D., Koenigsknecht, R.A., Malaney, G.D., Karras, J.E.: Factors related to doctoral dissertation topic selection. Research in Higher Education **30**(4), 357–373 (Aug 1989). https://doi.org/10.1007/BF00992560, https://doi.org/10.1007/BF00992560
13. Julian, K.: A phd should be about improving society, not chasing academic kudos. The Guardian (2018)
14. Keith, B., Vitasek, K., Manrodt, K., Kling, J.: Strategic sourcing in the new econ- omy: harnessing the potential of sourcing business models for modern procurement. Springer (2015)
15. Kuechler, B., Vaishnavi, V.: Promoting relevance in is research: An informing sys- tem for design science research. Informing science: The international journal of an emerging transdiscipline **14**(1), 125–138 (2011)

16. Larson, R.C., Ghaffarzadegan, N., Xue, Y.: Too many phd graduates or too few academic job openings: the basic reproductive number r0 in academia. Systems research and behavioral science **31**(6), 745–750 (2014)
17. Lei, S.A.: Strategies for finding and selecting an ideal thesis or dissertation topic: A review of literature. College Student Journal **43**(4), 1324–1333 (2009)
18. Lin, L., Green, C., Stamm, K., Christidis, P.: How long does it take to earn a research doctorate in psychology? Monitor on Psychology **48**(2)(2017)
19. McKenna, L., Gray, R.: The importance of ethics in research publications. Collegian **25**(2), 147–148 (2018). https://doi.org/10.1016/j.colegn.2018.02.006, https://doi.org/10.1016/j.colegn.2018.02.006
20. Norman, D.A., Verganti, R.: Incremental and radical innovation: Design research vs. technology and meaning change. Design issues **30**(1), 78–96 (2014)
21. Olalere, A.A., De Iulio, E., Aldarbag, A.M., Erdener, M.A.: The dissertation topic selection of doctoral students using dynamic network analysis. International Jour- nal of Doctoral Studies **9**(1), 85–107 (2014)
22. Olejniczak, T.: Rigour and relevance: A phd student's perspective. Manage- ment and Business Administration. Central Europe **23**, 113–130 (12 2015). https://doi.org/10.7206/mba.ce.2084-3356.160
23. Panda, A., Gupta, R.: Making academic research more relevant: A few suggestions. IIMB Management Review pp. 156–169 (09 2014). https://doi.org/10.1016/j.iimb.2014.07.008
24. Peters, R.L.: Getting what you came for: the smart student's guide to earning a master's or a Ph. D. Farrar, Straus and Giroux (1997)
25. Poock, M.C., Love, P.G.: Factors influencing the program choice of doctoral stu- dents in higher education administration. Naspa Journal **38**(2), 203–223 (2001)
26. Ségol, G.: Choosing a dissertation topic: Additional pointers. College Student Jour- nal **48**(1), 108–113 (2014)
27. Taylor, M., Martin, B., Wilsdon, J., et al: The Scientific Century: securing our future prosperity. The Royal Society (2010)
28. Toffel, M.: Enhancing the practical relevance of research. Production and Opera- tions Management **25**, 1493–1505 (03 2016). https://doi.org/10.1111/poms.12558
29. Turner, J., Bikson, T., Lyytinen, K., Mathiassen, L., Orlikowski, W.: Relevance versus rigor in information systems research: an issue of quality p. 34 (03 1991)
30. Useem, B.: Choosing a dissertation topic. PS: Political Science & Politics **30**(2), 213–216 (1997)
31. Whitworth, B., Friedman, R.: Reinventing academic publishing online. part i: Rigor, relevance and practice. First Monday **14**(8) (2009)
32. Wolff, J.: A glut of phds who can't find academic jobs. The Guardian (2015)
33. Woolston, C.: Graduate survey: Uncertain futures. Nature **526**(7574), 597–600 (2015)
34. Woolston, C.: Graduate survey: a love-hurt relationship. Nature **550**(7677), 549– 552 (2017)

Diachronic analysis of Spanish PhDs, focusing on Industrial oriented Doctorates

Jordi Moreno[1], Àlvar Hernández[1] and Andrea Valenzuela[1]

[1] Universitat Pompeu Fabra, Master in Intelligent Interactive Systems
{jordi.moreno01, alvar.hernadez01, andrea.valenzuela01}@estudiant.upf.edu

Abstract. We focus on the evolution of doctoral studies in Spain in terms of area of knowledge, age and gender. Part of the analysis is focused on finding evidences of the implementation of the Industrial Doctorates in Spain despite the objective of synchronizing the educational structures of Europe towards a common organization. The analysis has been also addressed in comparison to France. Finally, we have also carried out the analysis of Catalonia where Industrial Doctorates were implemented per se in 2012. The considered datasets are the '2017 Doctoral Thesis Database of the Spanish General Secretariat of Universities' and the '2006 Survey on Human Resources in Science and Technology from INE'. Finally, the 'Industrial Doctorates from *Generalitat de Catalunya*' dataset has been considered for the analysis of Catalonia. The main conclusions stracted are that the interest of Spain in Industrial Doctorates has been implicitly growing over time as the most industry related fields of knowledge are getting more attention. Moreover, the field of *Computer Science* has emerged as a popular subfield covered by doctoral studies. The mean age range for finishing the dissertation has not changed over time and there is no gender bias regarding the most recent data. Finally, although we are far away with other european countries is this matter, the pioneer implementation of Industrial Doctorates in Catalonia point towards a great success.

Keywords: doctoral studies**,** doctorates in Spain, Industrial Doctorates, gender bias.

1 Introduction

For the economies based on knowledge, a high qualified population in terms of research and technological development represents an added value for the region itself [1]. The universities are in charge of the generation of human capital with this high education which consist on a mean of eight years of intense hard-work. Nowadays, people spend a mean of five years to obtain the bachelor and the master's position and then three more years with the doctoral dissertation [2]. This last step prepares the student not only for the academic world -e.i researching or teaching at the University- but also for the incorporation to the companies. As the universities are the main responsible of the production of doctorates, they must adequate the curriculum against the needs presented in the job market [3]. The revision of the

changes that Universities do in the doctorates' curriculum against new improvements has been always a matter of interest and discussion.

In 2000, a group of european universities accepted one of the lines of impact of the Bologna[1] remodeling against education and create a pilot program called *Tuning* which had the objective of synchronize the educational structures of Europe towards a common organization. Until that common agreement, the universities were organized within an internal structure marked by their countries and the validation and mobility between universities and countries was a tedious affair [4].

Spain, as a member of the EHEA, also reformulated its structure towards higher education according to the European guidelines. In terms of the doctoral studies, there were different announcements –Berlin (2003), Bergen (2005), London (2007), Leuven (2009), Budapest-Vienna (2010), Bucharest (2012) - where European leaders in Higher Education delimited the basic characteristics that a PhD program must conform in the European environment. The events organized by the European University Association (2003, 2005, 2007, 2016) were also considered in the spanish process of building the new organization [5]. One of the most noticeable changes introduced in Spain is the possibility of a training model that places the doctoral student at the center of research in R&D&i projects, enabling quality, innovation, mobility and the internationalization of researchers. These new improvements are established in collaboration with the industry [3]. This new option of doctorate is called *Industrial Doctorate* and nowadays it is the only existing alternative to the traditional *Academic Doctorates* [1].

Regarding doctoral studies as a whole, we can confirm that Spain produces doctorates with the same rate as our surrounding countries. The taxes of doctorates doing research in Spain are also comparable to the taxes of the other countries [2]. Nevertheless, the rate of doctorates employed in companies (private sector) in Spain are approximately of the 50% with respect to the surrounding european countries. This noticeable difference has produced the increasing breach on the innovation boost that the industrial doctorates are trying to repair [2]. Apart from this fundamental problem, the industrial doctorates have also appeared due to the increasing demand of the technological companies [1].

In national terms, Madrid, Catalonia and Andalusia stand out. In 2011, these three regions contained the 56,2% of the doctorates of the whole country. Catalonia is a pioneer regarding the remodeling task of the doctorates' curriculum against the

[1] **Bologna:** The Bologna declaration lays the foundations of the European Higher Education Area (EHEA), organized according to certain principles (quality, mobility, diversity and competitiveness) and oriented towards the formation of an international network among the member countries that enables the development of common actions.

european tendency. It is the first Autonomous Community of Spain that offers doctoral dissertations more company-oriented, that is, industrial doctorates [2].

Although the steps performed during the alignment towards Europe has been fully documented, it is a slow process that is still ongoing. That is why the available information is spread in the Internet and there are not continuous statistical analysis of the whole process and its impact. Also papers analyzing Spanish PhD evolution do not consider the most recent data. Therefore, there are several questions that are still opened: How doctoral studies have evolved along time in Spain? Which are the most common areas of knowledge covered by doctorates? Is there a gender bias in doctorates? Do the statistics reveal the necessity of industrial doctorates? Which are the factors and the necessities that have led to the development of industrial doctorates? How do compare industrial related doctoral fields from Spain to those from some nearby countries? How the pioneer program in Catalonia has been evolving since its implementation in 2012 to actuality? All these questions will be addressed in the following pages.

2 Research methodology

This project is based on a brief literature review to understand the causes of the changes that have taken place in doctoral studies in Spain and also their evolution along time. In order to perform a more accurate analysis of the evolution and also of the characteristics of doctoral studies, two databases extracted from the spanish National Institute of Statistics (INE) and the Spanish Ministry of Education and Vocational Training, have been statistically analyzed. In concrete, the databases corresponding to 2006 [6] and 2017 [7] respectively, which have allowed us to analyse the areas of knowledge covered by doctoral studies, the ages of the doctorates once they finish the dissertation and if there is a gender bias regarding this stage of higher education. By comparing both databases we have also been able to explore the evolution of the target aspects under analysis and to identify the necessity of industrial doctorates. Findings of the statistical analysis have also been compared with the tendency of France as a neighbour and more advanced country in the matter. Finally, a third dataset has been used to concretely analyse the trajectory of the industrial doctorates in Catalonia since the catalan region is considered a pioneer in Industrial Doctorates. The database has been obtained from the catalan government [8] and it contains information of the number of projects offered per year, the type of projects covered as well as the type of companies that are participating. Moreover, it also contains information of the gender bias. All those aspects have been documented from 2012, where the pilot program was initiated, to 2018.

2.1 Search Strategy

In order to enclose the existing literature on the topic, the following keyword set has been defined for searching in Google Scholar, both in English and Spanish:

{industrial doctorate, professional doctorate, industrial doctorate in Spain, industrial doctorate Europe, doctorate business world, gender bias doctorate, satisfaction industrial doctorate}

Before analyzing the top papers of each entry, the most relevant papers were selected to be used as the main sources for this article. The selected papers were the ones that discussed the evolution of the doctorates in both Spain and Europe regarding our target topics. The number of citations and the impact factor of each article on their respective area of study have been also considered. In reading these, some other interesting sources were found in their references.

Finally, in order to complete the research, other information sources have been considered, such as websites of governmental organisations that have been double checked in order to assure their officiality and that will be fully referenced during the incoming sections.

3. Results

The analysis of results is divided in three sections. In *Section 3.1*, we assess the evolution of doctorates in Spain, considering 2006 and 2017 databases. In *Section 3.2*, we compare PhD students characteristics in the specific field of Engineering, between Spain in 2017 and France in 1996. And finally, in *Section 3.3* we examine the current performance of industrial doctorates in Catalonia since it recent emergence.

3.1 Spanish evolution considering 2006 and 2017 datasets

In order to evaluate the evolution doctoral studies in Spain, we are going to analyze the dataset 2017, Doctoral Thesis Database (TESEO). General Secretariat of Universities [7]. The total number of PhDs thesis in this database is 17,286. Firstly, we are going to examine which have been the changes in trends in spanish doctorates, focusing on some aspects of the students which have performed doctoral studies in a spanish University. To detect this changes, the just mentioned dataset from 2017 is going to be compared against the findings exposed in *J. F. Canal Domínguez et al.* [9], and also to the explicit database 2006, Survey on Human Resources in Science and Technology. (INE) [6] used in the same paper.

To be consistent with the methodology presented by *J. F. Canal Domínguez et al.* [9] and to perform an adequate comparison analysis of doctorates in Spain, we are going to classify the set of regulated PhDs into two main fields, taking into account their area of knowledge. These fields are *Social Sciences and Humanities* and *Sciences and Engineering*. As we can see in the appendix tables (from *Table 1* to *Table 5*), this categorization of the doctoral studies will comprehend rather the areas of knowledge present in the 2017, Doctoral Thesis Database (TESEO). General Secretariat of Universities or the ones in the 2006, Survey on Human Resources in

Science and Technology. (INE). The main subfields for *Social Sciences and Humanities* field will include *Social Sciences, Legal Sciences*, *Arts and Humanities*, and the *Sciences and Engineering* field consider *Engineering, Industry and Construction, Architecture*, *Agriculture, livestock, forestry, fishing and veterinary, Health Sciences* and *Natural Sciences*.

3.1.1 Popularity of area of knowledge

Examining the values of the percentage of doctors in each area of knowledge available at *Table 2* of the Appendix, one can notice that in 2006, there is a greater interest in the development of PhDs related with the *Science and Engineering* field (65.07% of the doctors) than with the *Social Sciences and Humanities* field (34.93% of the doctors). Another interesting observation is that the most popular subfield in the area of knowledge of *Sciences and Engineering* is *Natural Sciences* with a 29.45% of the total of doctors followed by *Health Sciences* with a 22.68%. The information in the database of 2017, available in *Table 1* of the Appendix, shows similar percentages as the ones presented in 2006 but, in this case, for the number of approved doctoral thesis. There are some interesting particularities in the data: The number of doctorates in *Science and Engineering* is again greater than the number of *Social Sciences and Humanities* doctorates. Nevertheless, the percentage has been slightly reduced in a 2.56% compared to the data of 2006. Also, it is significant the appearance of a new subfield of knowledge in 2017 which is *Computer Science*. If we consider *Computer Science* as part of the *Engineering and Technology* subfield -as it is commonly considered in Spain- and, therefore, we add up the two percentages, we have a total of a 12.24% of PhDs belonging to this new construct of the *Engineering and Technology* subfield.

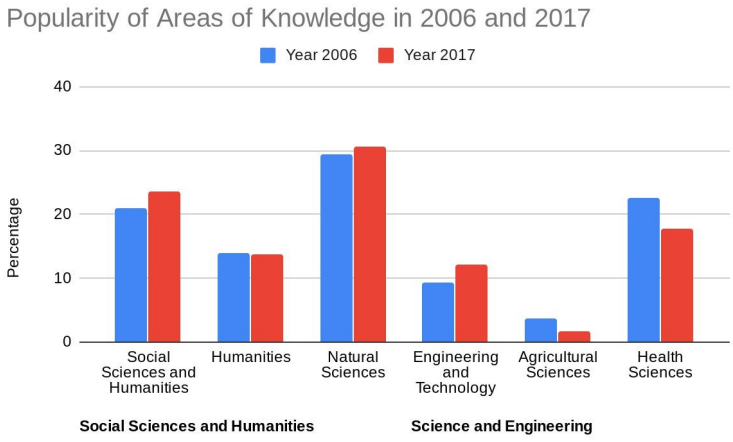


Figure 1: Percentage of doctorates in 2006 compared to approved PhD thesis in 2017 by field of knowledge.

As we can visually notice at *Figure 1*, this percentage of PhD developed in *Engineering and Technology* subfield for 2017 is higher than the one for doctors in the same subfield in 2006 (9.24%). Regarding this numerical evidences, we can observe that the areas of knowledge that are tightly related with the industry -such as *Engineering* and *Natural Sciences*- have increased in relevance along time and in comparison to the other areas of knowledge. Nevertheless, the reason for the aforementioned decrease of the 2.56% in the percentage of doctors in *Science and Engineering* field when comparing the data of 2006 with the data of 2017 is due to the reduction in the percentage of *Agricultural Sciences* subfield and in *Health Sciences and Social Services* subfield. As we expected, the areas which have gained importance are those which are most industry oriented. We consider them to be more industry oriented since they are most frequently selected as a subject for an Industrial Doctorate as we observe later on in *Section 3.3*.

3.1.2 Gender by area of knowledge

Regarding gender in the different areas of knowledge, *Table 4* of the Appendix shows the percentage of doctors of each field of doctorate and gender in 2006. The most perceptible fact is that the percentage of female doctors (45.77%) is slightly smaller than the percentage of male doctors (54.23%). Also, both female and male follow the general trend commented in *Section 3.1.1* since a greater amount of doctors choose the field of *Science and Engineering* instead of the *Social Sciences and Humanities* field. The subfield chosen by most of the doctors of both genders is *Natural Sciences*. *Figure 2* represents this gender differences of the data of2006.

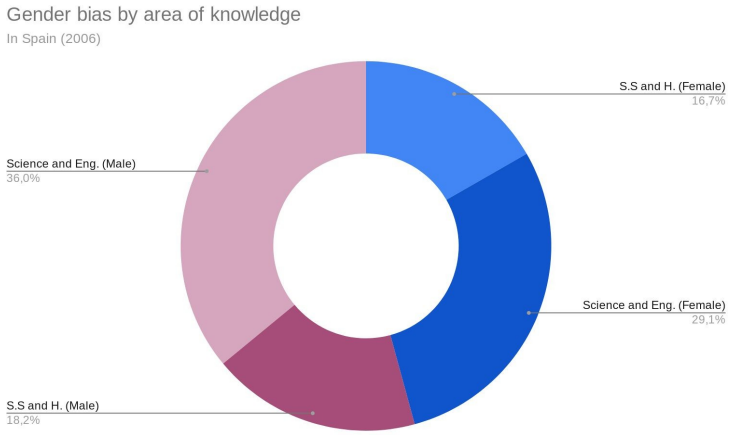


Figure 2: Gender bias by area of knowledge in Spain (2006). There is a subdivision regarding the most popular fields: *Social Science and Humanities* (S.S and H) and *Science and Engineering* (Science and Eng.)

Regarding the 2017's dataset, *Table 3* of the Appendix presents the number of approved doctoral thesis by gender that year. In this case, as the reader could infer from *Figure 3*, the percentage is higher for female than for male. Specifically, a 52.62% of female has been documented in contrast to a 47.38% of male. With respect to the proportion of thesis in the two main fields of knowledge, one can observe the same tendency presented in 2006 for both genders since there is also a greater percentage in the *Science and Engineering* field in comparison with the *Social Science and Humanities* field.

The datasets under analysis are difficult to compare due to the fact that they do not represent the same exact data. As commented before, the dataset of 2006 considers the number of doctorates whose thesis were being developed at that time while the database of 2017 collects the number of approved doctorates in 2017. Because of that, in this section, one cannot extract direct information about the evolution of the gender bias in time. Nevertheless, regarding the most recent data of 2017 (*Figure 3*), one could infer that there is not a significant bias regarding doctoral studies in general. In concrete, there are more females finishing their doctoral studies than males but this difference is nearly insignificant. Moreover, this difference is also small when comparing the two main areas of knowledge.

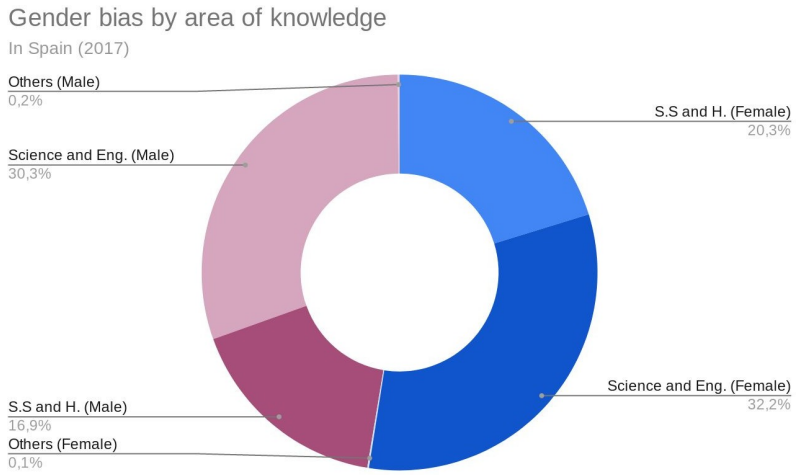


Figure 3: Gender bias by area of knowledge in Spain (2017).

3.1.3 Age by area of knowledge

As part of our analysis of the evolution of doctoral studies in Spain, this section is dedicated to drawing the tendency of the ages in which students successfully finished their PhD thesis in each area of knowledge. For that purpose, *Figure 4* shows the corresponding percentage of finished thesis by area of knowledge and range of age in 2017. It is interesting to note that, for all three considered areas of knowledge, the greatest percentage is achieved at the same age range: between 30 to 34 years. However, there are some differences when analysing in detail the different areas of knowledge. In the *Social Sciences and Humanities* field, there is a tendency towards relatively great percentages in age ranges above the most common one of 30-to-34. This field has a greater amount of elderly people finishing doctoral studies in comparison to the other fields. Therefore, we notice how, for greater age ranges, *Social Sciences and Humanities* field surpases *Sciences and Engineering* figures. In contrast, for the *Science and Engineering* subfields, the opposite tendency is observed. There is a very high percentage for the 30-to-34 age range while greater age ranges have low percentages. In this case, the observed tendency is that people that do not finish the doctorate in the range from 30-to-34 tend to finished sooner (24-to-29) than later (35-to-39 or above).

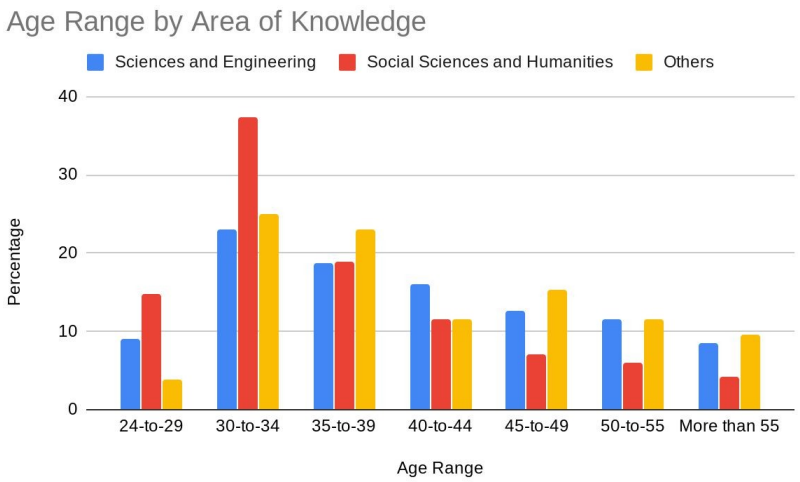


Figure 4: Age range by area of knowledge in Spain (2017). Based on *Table5* of the appendix.

J. F. Canal Domínguez et al. commented that, for doctors in 2006, an extra time of the 18% was required in the *Humanities* branch because of the different

characteristics of the doctoral studies depending on the field of knowledge [9]. We can conclude that this tendency has been maintained over the years.

3.2 Comparison of Spain and France for Engineering field

The results in *Section 3.1.1* point out an increase in doctoral subfields more industry-oriented, such as *Engineering* and *Computer Science*. This statistical results combined with the idea that industries are showing an increasing interest in doctorates [10], gives room to the introduction of the Industrial Doctorates as a new option for continuing higher education. The industries' interest has increased due to the fact that having a high qualified person carrying out a technological dissertation in a company facilitates an easy movement of the tacit knowledge [11]. Moreover, one should also consider the explicit changes performed in doctoral studies in order to align the spanish curriculum to the best practices promoted by the European Community which is undoubtedly wagering to Industrial Doctorates [4].

In this section, we are going to analyse the specific case proposed by *V. Mangematin* of 400 engineering science PhD students who graduated from the *Institut National Polytechnique de Grenoble*, between 1984 and 1996 [12]. The goal of this section is to find out the differences in between Spain and other nearby european countries -such as France- which has been encouraging the relationship between PhD students and industry a long before Spain [13].

In this concrete case, one could observe a higher propension of french engineering PhD students to collaborate with the private industry. Almost 50% of students have an industrial partner [12]. This piece of information should make the reader realize of how soon Industrial Doctorates started to be a reality in the neighbouring country [13].

With respect to the other target elements under analysis in this paper, the french dataset shows that three quarters of the doctorates were males. Regarding our analysis of gender bias and remembering the results found in *Section 3.1.2*, it is interesting to note that, in Spain, there is a higher diversity in engineering PhDs. Nevertheless, this extrapolation is not clear since other contextual factors should be taken into account when comparing two datasets which are quite far away in time. In terms of the age, we can also observe that the age in which the doctorate is completed is greater in the 2017 spanish sample. As shown in *Section 3.1.3*, the most common age range goes from 30 to 34 years old, while for the french sample the average age when PhD is completed is 28. In [9], the authors express that PhD thesis always take longer for those students who also work at the same time, which is a common practice in spanish doctorates. Therefore, there is an added difficulty in combining research training and work that may cause the necessity of expending more time on finishing the thesis. As with the gender analysis, no clear extrapolations should be performed by comparing this two datasets.

3.3 Catalonia as a pioneer region in Industrial Doctorates

There is not explicit data of the evolution of Industrial Doctorates in Spain since this type of doctoral studies has not been implemented in the whole country. The most complete dataset in term of doctoral studies as a whole is the one from 2017 that we have been considering along the entire paper. The considered dataset [8] has allowed us to draw several arguments in favor of the Industrial Doctorates but it does not consider this type of doctorates as an entity.

However, the catalan government published a first dataset with the evolution of the Industrial Doctorates per se. Catalonia is considered the most advanced Autonomous Community regarding the matter under study. The pioneer initiative in Catalonia was born in response to the challenge of transferring the leading technology and knowledge of global impact that our university and research system generates to the industries under the premise that all these contributions will reverse into economic and social development for the region [8].

On a global scale, Catalonia adopted this new modality of doctoral courses following the intention of the whole country to align their educational structures to the common european guideline. This concrete implementation is based on consolidated international experiences, such as the *Industrielles Conventions de la Formation per la Recherche* (CIFRE) in France or the *Industrial PhD Program* in Denmark, and is in line with other programs created subsequently, such as the European Industrial Doctorates (EID) of the European Commission [10]. As *Figure 5* shows, this kind of programs have shown an increasing interest and demand since their pilot implementation in 2012. In this last year (2018), a total amount of 98 projects has been offered by the catalan institutions.

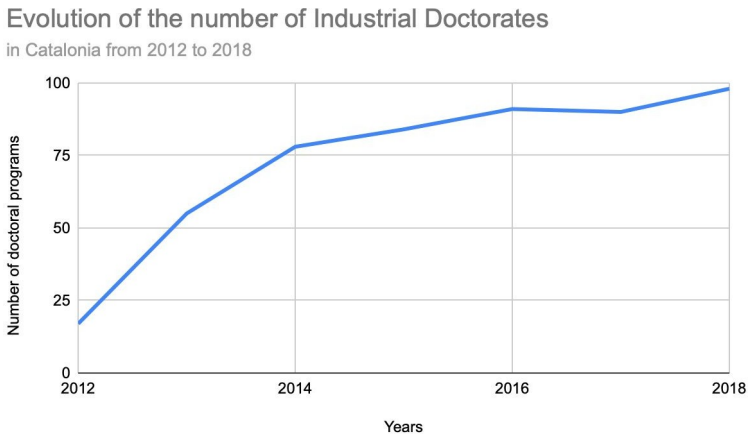


Figure 5: Evolution of the number of Industrial Doctorates in Catalonia from 2012 to 2018

Companies are interested in these types of projects because they allow them to bring people with knowledge and skills of high value, which they will use according to their needs and projects [10, 11]. Moreover, they will also have in their company potential leaders in innovation and research with a strong background. *Figure 6* shows the type of companies interested in this collaborations with the catalan universities in 2018. Small and Medium-Sized Enterprises (SMEs) represents the 55.2% of projects with respect to other types of companies due to the fact that they are an important driver of innovation within industry [14, 15]. Therefore, they are nearer than the Big Companies to the academic environment.

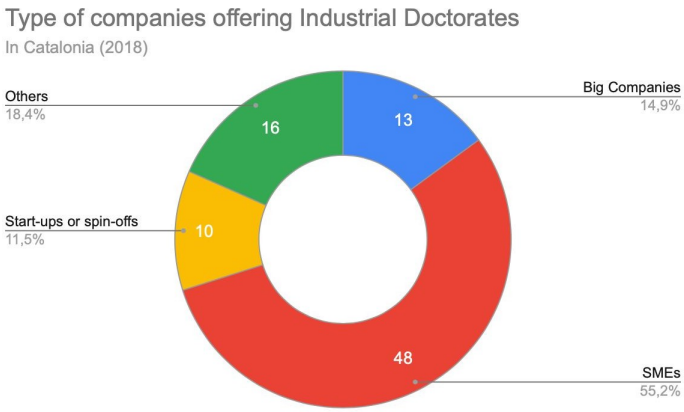


Figure 6: Type of companies offering Industrial Doctorates in Catalonia (2018)

3.4.1. Projects of the European Research Council. The most popular areas of knowledge.

The European Research Council (ERC) is a pan-European funding institution created to cover research and innovation in the European Union (EU). Catalonia has different industrial doctorates which are recipients of the ERC's grants [15]. In this framework, projects are divided into six categories: Social Sciences and Humanities (SH), Life Sciences (LS), Physics and Mathematics (PE1-3), Chemical sciences and materials (PE4-5), Information and communications technology (PE 6-7) and Product and process engineering and universe and earth system sciences (PE 8-10). *Figure 7* shows the distribution of the ERC projects from 2012 to 2018 [10].

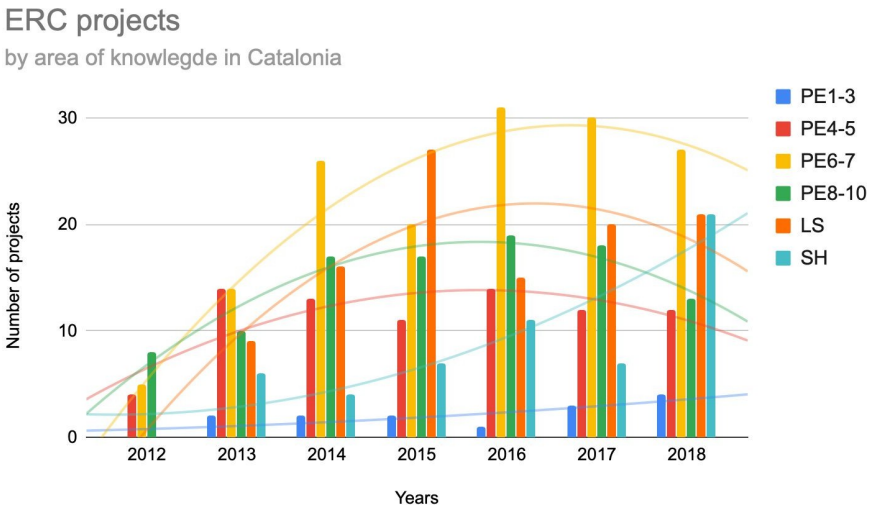


Figure 7: ERC projects by area of knowledge in Catalonia

As the reader could infer from *Figure 7*, the most popular topic within this last years corresponds to the area of Information and communication technologies. This result was expected since, as we have analyzed in the global case of Spain, the field of *Computer Science* has emerged due to the increasing applications in companies and their strong demand.

The second position in the most popular topics regarding industrial doctorates is occupied by the field of *Life Science* followed by the field of *Product and process engineering and universe and earth system sciences*. It is also interesting to notice the sharply increase of the industrial doctorates in Social Sciences and Humanities which could not be inferred when looking at the spanish dataset of2017.

Finally, and supporting our hypothesis that the appearance of the Industrial Doctorates are related with the increasing demand of the technological projects, one can easily observe how the more theoretical projects of the field of *Physics and Mathematics* have been maintained constant over the years. The more academic and theoretical projects have less demand when considering industrial projects.

3.4.2. Gender bias

Finally, as discussed when considering the evolution of the doctoral studies in Spain, this section analyses if there is a gender bias in the catalan industrial doctorates from 2012 to 2018.

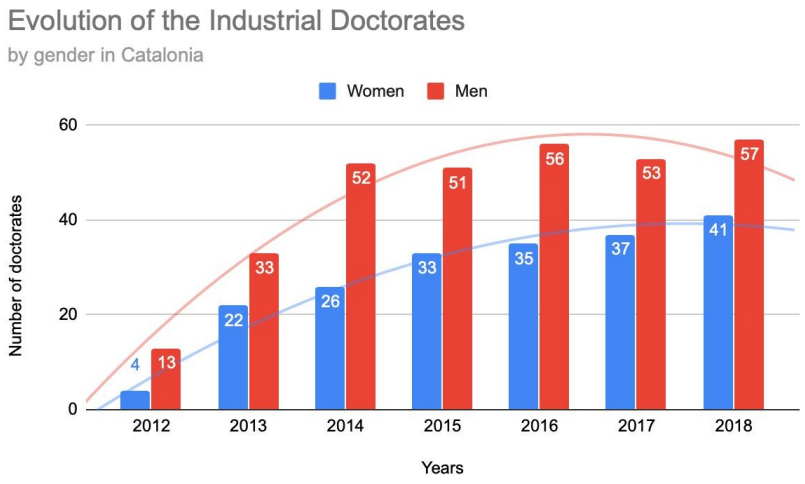


Figure 8: Evolution of the Industrial Doctorates by gender in Catalonia.

As *Figure 8* shows, the number of males doing an Industrial Doctorate has maintained constant during the last years (from 2014 to 2018) around a mean of 54 projects developed by males. Nevertheless, the number of females has increased during those years. This evolution is consistent since the number of projects has increased along time and the number of projects developed by males have maintained constant. Then, we can conclude this section by saying that nowadays there is gender bias in favour to male doctorands but the number of female doctorands is increasing as time goes by and it seems that they will overcome this difference in some years. In comparison with the spanish absence of gender bias regarding the data of 2017, the catalan evolution shows a clear tendency towards the gender equality in this higher education such as the one that Spain has achieved.

4. Conclusions

In this paper, we have addressed the analysis of the doctorates in Spain focusing on the evolution of three target aspects -area of knowledge, age range and gender- and, from the first target aspect, we have tried to infer the reasons for the emergence of Industrial Doctorates. Moreover, we have analyzed the position of Spain with respect to France as a nearby european country and the evolution of the most pioneer spanish region in the matter. From here, several conclusions could be extracted also answering to our initial research questions proposed at the beginning of the article.

Regarding the evolution in terms of the areas of knowledge, we have observed the introduction of a new subfield of *Computer Science* in the field of *Science and Engineering.* Although the popularity of the whole field has slightly decreased since 2006, the most industry-related subfields -such as *Engineering and Computer Science*- have suffered an increase in the number of graduates pointing towards the direction of Industrial Doctorates. In terms of the mean age of a person who finishes the dissertation, the most common mean age ranges from 30-to-34 and it is noticed that people doing their dissertation on the *Sciences and Engineering* field tend to be younger than those in the *Social Sciences and Humanities* field. Similar results are observed for both 2006 and 2017 cases. Finally, in terms of the gender bias, we were not able to draw relevant conclusions. Nevertheless, regarding the latest dataset of 2017, we have observed that the gender bias of the total amount of doctoral studies is small. Both female and male categories are balanced.

Comparing the current situation of Spain with France data in 1996, we have realized that other nearby countries started to wager for company-oriented projects a long time before Spain. As pointed out when analysing the data, no further conclusions could be extracted by comparison of time-distant datasets since there must be more contextual aspects that should be considered.

Finally, in the study of Catalonia as a pioneer region in Industrial Doctorates in Spain, we can firstly conclude that the interest in this kind of PhDs have been growing since their first implementation in 2012. The type of companies more interested in offering Industrial Doctorates are Small and Medium-Sized Enterprises which are the nearest to the academic environment. The most relevant topic in industrial doctorates is *Information and Communication Technologies*, related to the *Computer Science* field of knowledge, which, as commented before, tends to be recurrent in industry projects. Finally, we note that the number of female doctors in Industrial Doctorates is lower than male doctors. In the case of Catalonia, a significant gender bias has been found. Nevertheless, the data shows a clear tendency towards a balance as time goes by.

With this, we can conclude that the interest in Spain in the currently so-called Industrial Doctorates has been implicitly growing over time, by acquiring more attention the most industry related fields of knowledge. Moreover, the explicit result of this trend can be shown regarding the situation of Catalonia where, because of the spanish process of matching the european higher education normative, has implemented Industrial Doctorates per se and they have shown a clear success and demand.

4.1 Future work

For continuing this study it could be interesting to analyse all the regions in Spain that have implemented the option of Industrial Doctorates in order to see if the particular case of Catalonia represents the general trend of the country. Also, if other national and reliable datasets containing more modern data, i.e 2018 and 2029, were available, we could address the time evolution of the doctoral studies in the past three years and give a more accurate analysis of the current state of the three target elements under study (area of knowledge, age rate and gender) and its recent evolution.

Finally, another interesting factor could be the analysis of where doctorates work after their Academic or Industrial PhDs in order to analyse which is the most general trend and also if this new option of Industrial Doctorates makes sense regarding type of job after the dissertation and also doctorate/job satisfaction. Nevertheless, this last point is more out of the scope of our initial purpose.

5. References

[1] Bonito, M. B. & Ayllón, R. R. La aportación de los doctores al desarrollo económico y social a través de su contribución a la I+D+i. Estudios CYD 05/2014.

[2] Bonito, M. B., Torrubias, P. G., & Ayllón, R. R. El empleo de los doctores en España y su relación con la I+D+i y los estudios de doctorado. Conferencia de Consejos Sociales de las Universidades Públicas Españolas. Colección estudios e informes (2014).

[3] Jiménez-Ramírez, M. Los nuevos estudios de doctorado en España: avances y retos para su convergencia con Europa. Revista Iberoamericana de Educación Superior. No21, Vol VIII, pp. 123-137, (2017).

[4] European Higher Education Area and Bologna Process. Retrieved December 10, 2019, from http://www.ehea.info

[5] Coordinadora de Representantes de Estudiantes de Universidades Públicas. Informe marco del EES. Espacio Europeo de Educación Superior: Construcción y Seguimiento. Online available at (Last search: 09/12/2019): http://dge.ugr.es/delegacion/pages/_creup/_informescreup/informe-marco-del-eess-construccion-y-seguimiento/

[6] Spanish National Institute of Statistics (INE). (2008, July 11). Survey on Human Resources in Science and Technology. 2006. Retrieved December 10, 2019, from https://www.ine.es/dynt3/inebase/index.htm?type=pcaxis&path=/t14/p225/a2006&file=pcaxis

[7] Doctoral Thesis Database (TESEO). General Secretariat of Universities. (2018, October 30). Doctoral Thesis Statistics (DTE). Doctoral thesis approved. 2017. Retrieved December 10, 2019, from http://estadisticas.mecd.gob.es/EducaDynPx/educabase/index.htm?type=pcaxis&path=/Universitaria/Tesis/2017/&file=pcaxis#
[8] Generalitat de Catalunya. (n.d.). Doctorats industrials. Estadístiques. Retrieved December 10, 2019, from http://doctoratsindustrials.gencat.cat/ca/contents/view/23
[9] Canal Domínguez, J. F., & Muñiz Pérez, M. A. (2012). Professional Doctorates and Careers: the Spanish case 1. *European Journal of Education, 47*(1), 153-171.
[10] Sauermann H. & Roach M. Not all scientists pay to be scientists: PhDs' preferences for publishing in industrial employment. Research Policy. vol: 43 (1) pp: 32-47 (2014).
[11] Stephan P.E, Sumell A.J., Adams J.D. & Black G.C. Firm placements of new PhDs: Implications for Knowledge Transfer. The role of labour mobility and informal networks for knowledge transfer. Chapter 7.
[12] Mangematin, V. (2000). PhD job market: professional trajectories and incentives during the PhD. *Research policy*, *29*(6), 741-756.
[13] Heraud, J. A., & Levy, R. (2005). University-industry relationships and regional innovation systems: Analysis of the French procedure cifre. In *Innovation Policy in a Knowledge-Based Economy* (pp. 193-219). Springer, Berlin, Heidelberg.
[14] Mrva, M., & Stachová, P. (2014). Regional development and support of SMEs–how university project can help. Procedia-Social and Behavioral Sciences, 110, 617-626.
[15] Gunasekaran, A., Putnik, G. D., Peças, P., & Henriques, E. (2006). Best practices of collaboration between university and industrial SMEs. Benchmarking: An International Journal.

6. Appendix

Knowledge Area		Number	Percentage	Group percentage
Social Sciences and Humanities	**Social Sciences, Journalism and Documentation**	1 950	11.56%	37.17%
	Business, Administration and Law	1 069	6.34%	
	Arts and Humanities	2 305	13.67%	
	Education	945	5.6%	
Science and Engineering	**Sciences**	5 177	30.7%	62.51%
	Computer Science	822	4.87%	
	Engineering, Industry and Construction	1 243	7.37%	
	Agriculture, livestock, forestry, fishing, and veterinary	288	1.71%	
	Health and social services	3 012	17.86%	
Others	**Services**	52	0.31%	0.31%

Table 1: Number of Doctoral thesis approved by the area of knowledge. Year 2017

Knowledge Area		Percentage	Group percentage
Social Sciences and Humanities	**Social Sciences**	21.02%	34.93%
	Humanities	13.91%	
Science and Engineering	**Natural Sciences**	29.45%	65.07%
	Engineering and Technology	9.24%	
	Agricultural Sciences	3.70%	
	Health Sciences	22.68%	

Table 2: Percentage of doctorates by branch of education. Year 2006

Knowledge Area		Female	Female Percentage per Group	Male	Male Percentage per Group
Social Sciences and Humanities	**Social Sciences, Journalism and Documentation**	1 077	20.25%	873	16.92%
	Business, Administration and Law	494		575	
	Arts and Humanities	1 232		1 073	
	Education	612		333	
Science and Engineering	**Sciences**	2 689	32.23%	2 488	30.29%
	Computer Science	213		609	
	Engineering, Industry and Construction	469		774	
	Agriculture, livestock, forestry, fishing, and veterinary	131		157	
	Health and social services	1 933		1 079	
Others	**Services**	23	0.14%	29	0.17%
Total Percentage			52.62%		47.38%

Table 3: Number of doctoral thesis approved by area of knowledge and sex. Year 2017

Knowledge Area		Female	Female Percentage per Group	Male	Male Percentage per Group
Social Sciences and Humanities	**Social Sciences**	9,80%	16.69%	11.22%	18.24%
	Humanities	6.89%		7.02%	
Science and Engineering	**Natural Sciences**	13.51%	29.08%	15.94%	35.99%
	Engineering and Technology	2.54%		6.70%	
	Agricultural Sciences	1.71%		1.99%	
	Health Sciences	11.32%		11.36%	
Total Percentage			45.77%		54.23%

Table 4: Percentage of Doctors by field of Doctorate and sex. Year 2006

	24 to 29 years	30 to 34 years	35 to 39 years	40 to 44 years	45 to 49 years	50 to 55 years	More than 55 years
Education	7.41	19.37	18.52	16.19	13.76	13.65	11.11
Arts and Humanities	11.24	26.28	18	14.92	9,41	9.07	8.07
Social Sciences, Journalism and Documentation	9.33	26.36	21.85	15.44	11.23	9.74	6.05
Business, Administration and Law	7.86	20.02	16.18	17.4	15.9	13.84	8.79
Sciences	24.07	44.47	15.65	7.13	3.84	2.9	1.95
Computer Science	14.72	43.8	17.88	12.04	6.45	3.41	1.7
Engineering, Industry and Construction	11.67	31.94	19.79	13.03	9.25	8.85	5.47
Agriculture, livestock, forestry, fishing, and veterinary	15.28	36.11	17.36	10.42	7.29	7.29	6.25
Health and social services	7.84	30.74	23.94	15.01	8.76	7.93	5.78
Services	3.85	25	23.08	11.54	15.38	11.54	9.62

Table 5: Percentage of approved doctoral thesis by field of study of thesis and age group. Year 2017

Evidence for a mental health crisis in doctoral students

Georgios Angelopoulos [1], Daniel Levkovits[2], Jorge Pimienta[3], and Jonatan Koren[4]

Master in Intelligent Interactive Systems
Department of Information and Communication Technologies
Universitat Pompeu Fabra, Barcelona, Spain.
[1]georgios.angelopoulos01@estudiant.upf.edu
[2]danielsteven.levkovits01@estudiant.upf.edu
[3]jorge.pimienta01@estudiant.upf.edu
[4]ionathanaharon.koren01@estudiant.upf.edu

Abstract. In recent years, journalists, research policy observers, and academics have voiced concerns about the potential impact of research conditions in universities on mental health problems. These concerns are often related to recent shifts in the organization of academic research, such as increased workloads, intensification and the pace of change. The aim of the present study is to asses the prevalence of mental health, gender bias and social difficulties in Doctorates. For the above mentioned, this work collects valuable information from 80 doctoral students residing in Spain in a broad category of research areas using a questionnaire. Our results showed that more than 50% of Ph.D. students are at risk of having or developing a common psychiatric disorder. As doctoral students with mental health issues may pose a considerable cost to research institutions and teams, this work represents a call to action for the Ph.D. system.

Keywords: Mental health, PhD students, PhD difficulties, Doctoral Studies

1 Introduction

As most Ph.D. students are part of larger research teams, whose composition de- termines scientific impact, doctoral students with mental health issues may pose a considerable cost to research institutions and teams. To date, research policy efforts seemed to have focused more on ”hard outcomes” such as publications, impact factors, and patents, while ignoring the health effects of ”soft” policy outcomes, such as stress and anxiety. However, soft outcomes may create serious financial costs for research institutions. In addition, the mental health problems of Ph.D. students impact both the supply and entrance to the research indus- try. Organizational policies that are linked to mental health problems will lead individuals to quit their Ph.D. studies or leave the research industry altogether [1].

There has been an increment in studies focusing on the effects of doctoral stu- dents. In general, the widespread opinion about university students is that they are individuals with low-stress levels in environments that motivate recreation and diversity [2]. But, a great number of Ph.D. students have several difficulties to face through doctoral studies: lifestyle changes, customs, worldviews, socioeco- nomic status, language difficulties for international students, studies themselves and the research. In addition to the previously mentioned difficulties, studies have shown [3] social isolation plays an important role, students tend to feel socially unaccompanied during their studies.

University environment and research conditions have an impact on the social and psychological situation of the student. The study period to finish a doctoral study varies between 3 to 7 years, depending on the specific subject, university, and student. Over this period, the first year tend to be used to prepare the re- search subject for doctoral studies. Attrition rates among doctoral candidates have been reported to range between 30% and 50% [4]. This has a direct re- lationship to stress and isolation, feelings that were found to be the first and more influential contributors to the phenomenon of dissertation [5]. Despite the above mentioned, it has been an increase in Ph.D. eligibility, "OECD countries reported a 56% increase between 2000 and 2012" [6].

Researchers supervisors have a big impact on the students, they provide guidance during the studies. Supervisors can be either positive, when they sup- port and guide the student socially, psychologically and on educational life, or negative, when they ignore and not provide feedback along the way of the doc- toral study. On the other hand, research studies have also raised concerns about gender bias in doctoral studies, as many women claimed to be discriminated in terms of study space, freedom of action, supervisors evaluation, and examiners. Other difficulties arise in the context of employment, professional past affect- ing research, job offers during studies increase the dissertation rate and future employment possibility affects student research performance [5].

To sum up, given the potential importance of mental health problems for research policy, there is an urgent need for systematic empirical data rather than anecdotal information on their prevalence and the organizational policies that are linked to them. Given the current lack of an empirical basis for mental health concerns and solutions, the current study has three aims. First, we aim to inform research policy by assessing mental health prevalence in a representative sample of doctoral students in Spain. Second, to assess the scope of the problem, we compared the mental health of Ph.D. students with other university students. Third, with the aim of a better understanding of how research and organizational policies may relate to mental health, we examined doctoral students' perceptions of the academic environment and linked them to mental health problems.

The remainder of this paper is organized as follows. Section 2 presents the research methodology. In Section 3, the results are presented. Finally, Section 4 concludes our paper.

2 Research methodology

The research problem addressed by this study focuses on the experience of Ph.D. students and the possible negative emotions derived from it. The overall ap- proach of this research was to analyze the results of similar studies on this topic, which corresponds to a literature review, intending to identify the most critical and relevant elements of research from different sources and perspectives. More- over, these results were contrasted with current Ph.D. students by making use of a survey based on different categories. The survey was selected to be the main instrument for collecting the data, as it offered an anonymous channel between the parties.

2.1 Literature Review

During the literature review, it was noticeable that one of the main concerns of the researchers, was the **Psychological Impact** on the Ph.D. students. For instance, on the paper [7], it is studied the emotional labor of the students while conducting their studies and highlights the importance of monitoring the student's emotional status, their relationship with supervisors, their relationship with their colleagues, and the academia itself.

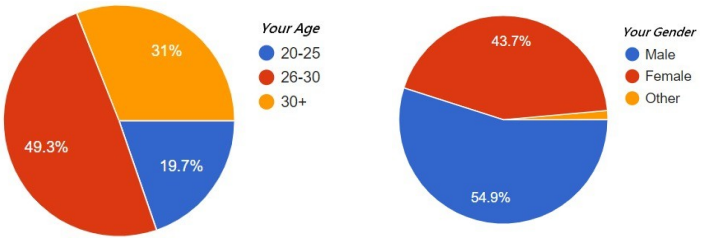
Also, the **Gender Bias** was an important topic of discussion among the re- searchers, this situation is addressed in [8] who analyze the relationship between some challenges and how they are related to the gender of the student. It pro- poses an analysis of several variables and situations that may create a bias during the Doctorate program; the expected role of a woman in society, the time com- mitment, female identity problems, family commitments, and others. Although the majority of the discussions are focused on one gender, for the data collection stage, this research will consider any type of gender bias or discrimination.

Finally, the **Structure of the Doctorate Itself**, including aspects like the schedule, the tutorship sessions, etc. were also a relevant factor to consider for this study, as mentioned in [9] in where it examined the importance of the role of the supervisor during the doctorate program, what are the expectations from the student side. It covers the traditional educational model and the select important variables, namely, Flexibility of the program, time dedication, and mobility.

3 Results

The research instrument used to extract more information was a questionnaire, it was designed so it collects information from doctoral students residing in Spain asking them 13 questions based on their Ph.D. daily life. The questions were selected from the elements mentioned in the previous section.

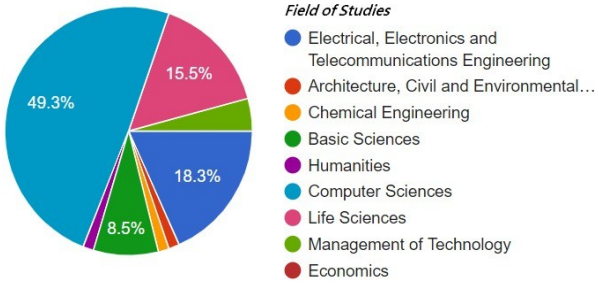
The questionnaire was administered online using a secure survey tool, Google Forms. This allowed us to ensure the anonymity of respondents but also to have accurate analysis from our data. By the closing date 80 responses had been received with similar numbers of male and female survey participants (see Fig. 1).



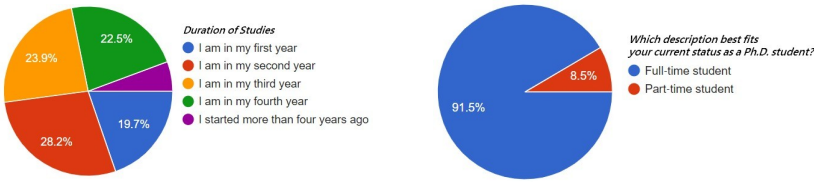
(a) Age (b) Gender

Fig. 1: Age and Gender of the Participants

The majority of participants belong to the Electrical, Electronics, and Telecom- munications Engineering, which represents 49.3% of the sample. This sample of students that answered the survey, were mostly in the middle stages of their re- spective students, from which 91.5% were participating in a full-time program. (see Fig. 2)



(a) Field



(b) Duration (c) Status

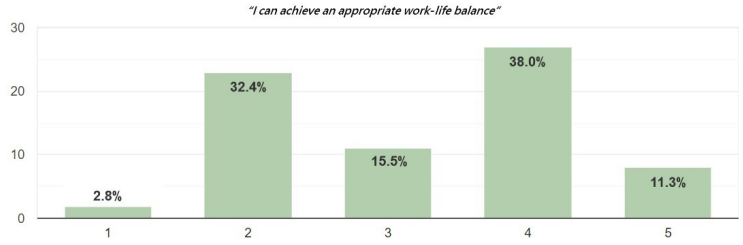
Fig. 2: Field, Duration and Status of studies

On Fig. 3, we asked our participants how regularly did they meet with their supervisor. The results show that 47.9% of the students answered about once per two weeks or less. Studies have shown that the Supervisors' support is central to the participants' stories; it is thus assumed to play a role in the psychology of

the student [10]. Frequency of meetings is also linked to the duration of studies, with the frequency of meetings declining between the first and fourth years.

Fig. 3: Relationship with the Supervisor

According to a report by the Organisation for Economic Co-operation (OECD) [11], Spain has one of the best rankings of work-life balance in a study of a group of the world's advanced economies, therefore this is reflected on the Fig. 4, which only 35.2% of our surveyees answered negatively in contrast with their quality of the studies whom most of them approve. This can, of course, derive on a dif- ferent research path, i.e. the domain of the program, or any other unintentional bias that could be generating this contradiction.



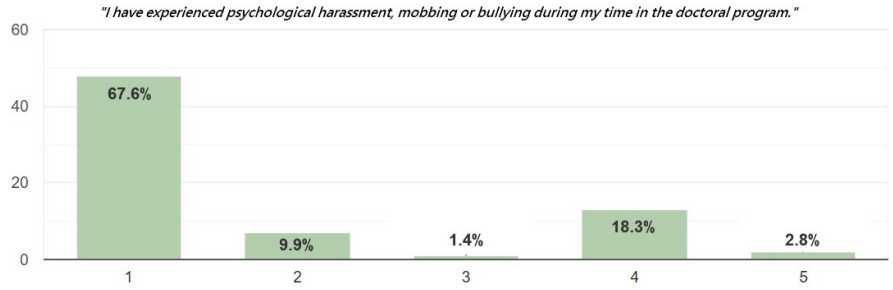
(a) Work-Life Balance

(b) Quality of PhD

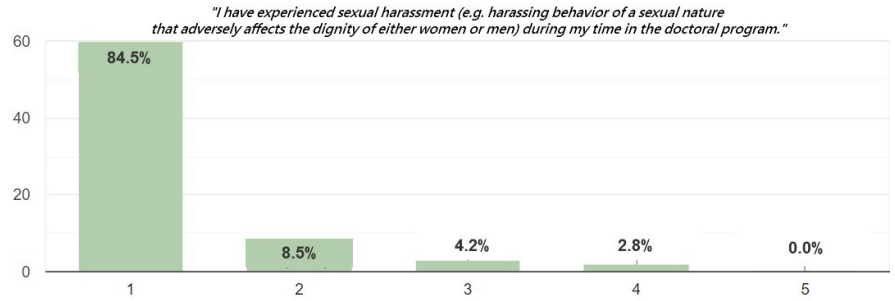
Fig. 4: Doctorate Satisfaction (1-Completely Disagree, 5-Completely Agree)

Two different ways of assessing negative workplace experiences, bullying and harassment were used. First, doctoral students were asked for their self-assessment of whether they had experienced bullying or sexual harassment (see Fig. 5). Using the self-report measure, it is evident that the vast majority of stu- dents indicate that they have not experienced sexual harassment. Nonetheless, it is notable that 2.8% indicate (i.e. 'agree' or 'completely agree') that they have been sexually harassed.

Although 77.5% of respondents indicate that they have not been bullied, over 1.4% are unsure as to whether the behavior they have experienced counts as psychological harassment, mobbing or bullying, and a further 21% indicate that they have experienced such behaviors (i.e. 'agree' or 'strongly agree').



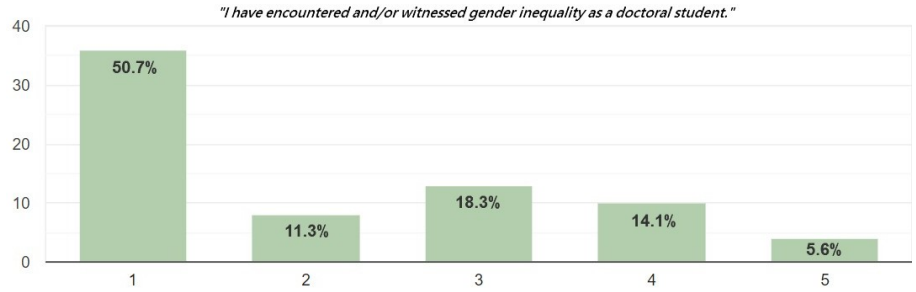
(a) Psychological Harassment



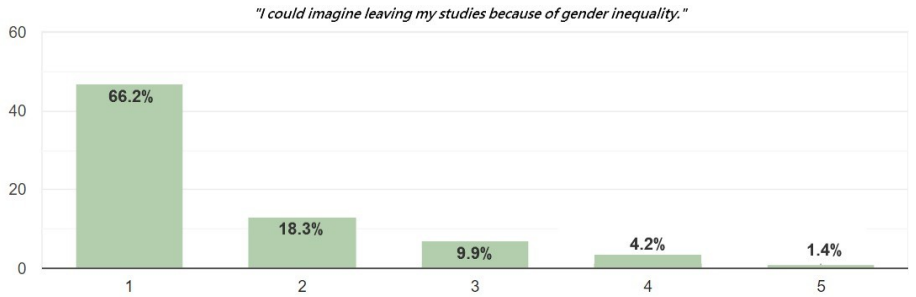
(b) Sexual Harassment

Fig. 5: Psychological and Sexual Harassment (1-Completely Disagree, 5-Completely Agree)

Regarding encountered gender inequality, and also considering the possibility of leaving the students because of incidents related to this aspect. it was found that around 40% of this sample have witnessed some type of Gender inequality, However, less than 15% will consider leaving their respective students due to that reason.



(a) Encountered/Witnessed Gender Inequality



(b) Leaving Studies

Fig. 6: Gender Inequality (1-Completely Disagree, 5-Completely Agree)

What is the connection between these work conditions and experiences and student mental health? Studies have shown [12] that workplace conditions have tremendous effects on mental health.

In order to obtain a general and more detailed description of the negative feel- ings experienced by the students, we requested the participants to select one or more from a list of Symptoms. We highlight the ones with the highest frequency (i.e. those with more than 50% of selection), in particular, the persistent feeling of sadness and Reduced ability to concentrate were on the top with 64.6%, and also Daily problems or stress, with 60%, which show that most of the challenges experienced during the program have a systematic affection of the emotions of the students. As mentioned by [13] which describes the doctoral experience with a journey and how students develop themselves through the whole career, the author compares the program with a quest in which the student must go through several challenges. It mentions the importance of the willingness to succeed in order to have a good performance, however, it seems not to agree with a 60% of students who experience a lack of motivation, which is worsened by a Significant tiredness or sleeping problems indicators of 58.5%.

Other aspects are not meant to be underestimated, on the contrary, they all support the same trend. For instance, considering that 7.7% of the participants

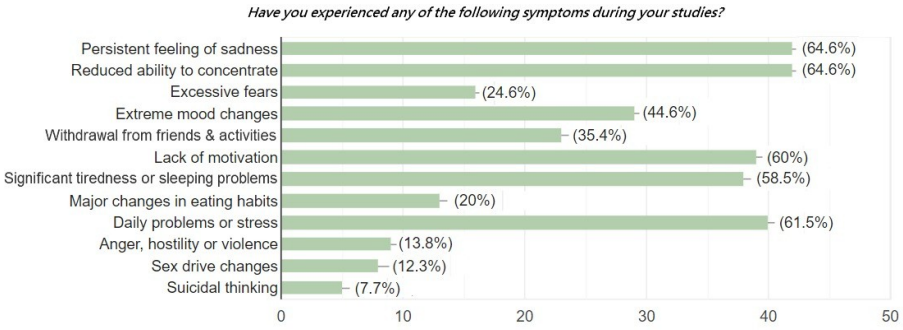


Fig. 7: Psychological Symptoms during studies

are experiencing suicidal thoughts, or 13.8% experiencing episodes of Anger, hostility or violence, encourage a further investigation on the causes and triggers. A study published in the American Journal of Orthopsychiatry in 2013 [14], looked at mental health in 1,700 university students, shows that the estimated prevalence of any depressive or anxiety disorder was 15.6% for undergraduates and 13.0% for graduate students. Suicidal thoughts were reported by 2% of students. In contrast, our results show that doctoral students are **signiftcantly more** likely to experience depression and anxiety as compared to the other students.

4 Conclusions

Resembling at the previous results, it's hard not to conclude that the current Ph.D. system is fundamentally unaware of the fact that mental health issues are rife: many doctorate students are at risk of having or developing a psychi- atric disorder like depression. The high prevalence of mental health problems in Ph.D. students is critical in terms of individual suffering, organizational and societal costs. In the long run, however, it will also impact on the research itself. Therefore, It is encouraged for the scientific community to further evaluate and disrupt the current Ph.D. system to make it better for early-career researchers. Confucius said one of the core principles of the academy should be as follows: "The essence of knowledge is, having it, to apply it". Reminding ourselves of this may help to fix the naive Ph.D. machine.

References

1. Podsakoff, N. P., Lepine, J. A. & Lepine, M. A. Differential challenge stressor-hindrance stressor relationships with job attitudes, turnover intentions, turnover, and withdrawal behavior: A meta-analysis. *Journal of Applied Psychology* **92**, 438–454 (2007). Cited By :672.

2. Bozeman, B. & Gaughan, M. Job satisfaction among university faculty: Individual, work, and institutional determinants. *Journal of Higher Education* **82**, 154–186 (2011). Cited By :82.
3. Haynes, C. *et al.* My world is not my doctoral program. . . or is it?: Female students' perceptions of well-being. *International Journal of Doctoral Studies* **7**, 1–17 (2012).
4. Pyhälto, K., Toom, A., Stubb, J. & Lonka, K. Challenges of becoming a scholar: A study of doctoral students' problems and well-being. *ISrn Education* **2012** (2012).
5. Jones, M. Issues in doctoral studies-forty years of journal discussion: Where have we been and where are we going? In *Proceedings of the Informing Science and Information Technology Education Conference*, 83–104 (Informing Science Institute, 2013).
6. Levecque, K., Anseel, F., De Beuckelaer, A., Van der Heyden, J. & Gisle, L. Work organization and mental health problems in phd students. *Research Policy* **46**, 868–879 (2017).
7. Nutov, L. & Hazzan, O. Feeling the doctorate: Is doctoral research that studies the emotional labor of doctoral students possible? *International Journal of Doctoral Studies* (2011).
8. Carter, S., Blumenstein, M. & Cook, C. Different for women? The challenges of doctoral studies. *Teaching in Higher Education* (2013).
9. Baptista, A. V. Challenges to doctoral research and supervision quality: A theoretical approach. *Procedia - Social and Behavioral Sciences* **15**, 3576–3581 (2011).
10. Devos, C. *et al.* Doctoral students' experiences leading to completion or attrition: a matter of sense, progress and distress. *European Journal of Psychology of Education* (2017).
11. Lncs homepage. http://www.oecdbetterlifeindex.org/topics/work-life-balance/s (2019). [Online; accessed 29 Nov 2019].
12. Woo, J.-M. & Postolache, T. The impact of work environment on mood disorders and suicide: Evidence and implications. *International journal on disability and human development : IJDHD* **7**, 185–200 (2008).
13. Hughes, C. & Tight, M. The metaphors we study by: The doctorate as a journey and/or as work. *Higher Education Research and Development* (2013).
14. Eisenberg, G. S. E. G. E., D. & Hefner, J. L. Prevalence and correlates of depression, anxiety, and suicidality among university students. *American Journal of Orthopsychiatry* 534–542 (2007).